Lessons From Using Genome Editing to Create a Faithful Model of a Novel Congenital Anemia

Lessons from Using Genome Editing to Create a Faithful Model of a Novel Congenital Anemia

A thesis presented by

Xiaoli Mi

to

Harvard Medical School
&
Harvard-MIT Division of Health Sciences and Technology (HST)

in partial fulfillment
of the requirements for the M.D. Degree
with Honors in a Special Field

Boston, Massachusetts
February 2018

This thesis is dedicated to my parents,
for their unconditional love.

ABSTRACT

The ease of genome sequencing over the past decade has greatly facilitated the identification of genetic mutations causing a large number of human diseases. However, in the vast majority of cases, how these mutations lead to disease remains an enigma. Targeted genome engineering via the RNA-guided CRISPR/Cas9 nuclease system offers the possibility of faithfully mimicking molecular mutations in their endogenous context to study underlying mechanisms. In this study, I explored the question of whether genome editing is an effective approach to model a novel genetic disease. Recently, our laboratory has identified two patients with a distinct form of congenital dyserythropoietic anemia, characterized by erythroid and megakaryocytic dyplasia and mild elevations in fetal hemoglobin levels. Curiously, the coding sequence of DNA from the patients was unremarkable but further analysis revealed a common hemizygous point mutation in the last intron of *GATA1* (chrX: 48,652,176 C>T in hg19). This gene encodes the hematopoietic master regulatory transcription factor GATA1 essential for the differentiation of erythrocytes and megkaryocytes. Genetic mapping and segregation within the families provided strong evidence to support a causal relationship between the mutation and disease. To model the disease in a tractable system and further understand the underlying basis by which the mutation results in disordered blood production, I performed genome editing with CRISPR/Cas9 in a human erythroid cell line to recreate this intronic *GATA1* mutation. I designed and cloned nine distinct single guide RNAs (sgRNAs) and identified one that cleaved the target locus in an efficient manner. Using this sgRNA, I showed that an increased mass ratio of exogenous repair template to sgRNA to Cas9 dramatically improved *GATA1* editing efficiency and diversity. The rate of genomic perturbation increased from 20% to 80% and incorporation of the intronic mutation via homology directed repair occurred at 3.8%. Using three computational

methods, I characterized all allele-specific editing events in 133 isogenic cell lines and mapped them by genomic position relative to the Cas9 cleavage site. I found that most *GATA1* changes occurred within 15 bp of the Cas9 cleavage site and the editing frequency is higher upstream than downstream at each position equidistant from the predicted double-strand breaks. Consistent with this observation, most *GATA1* edits occurred within the intron and those that altered the downstream canonical splice acceptor site and exon were exclusively heterozygous. These findings suggest that the human erythroid cell line disfavors alteration of sequences essential for proper function of GATA1, via splicing regulation or protein coding. Notably, the desired homozygous C>T mutations co-occurred with additional modifications such as deletion of an alternative splice acceptor site and the protospacer adjacent motif (PAM) required for Cas9 binding. It is possible that concurrent deletion of the alternative splice acceptor site rescues the splicing defect from the mutation and genome editing recurred until PAM was rendered unrecognizable. Following these experiments, my colleagues demonstrated that the C>T mutation results in activation of the alternative splice acceptor site and a partial intron retention event using transient expression of a minigene. The observed splicing alteration was confirmed in patient samples and further experiments showed that the resultant GATA1 protein is inactive. While many recent proof-of-principle studies have described successful application of genome editing to generate models of common diseases with well-established genetic alterations, no study has yet modeled a novel disease with a novel mutation and provided insight into disease pathophysiology. I conclude with important limitations of using CRISPR/Cas9 for disease modeling and discuss possible solutions and efficient alternatives. These lessons have broad implications for future applications of genome editing and classical approaches to study disease in an effort to generate functionally and clinically relevant knowledge.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTERS 1 & 2

LIST OF FIGURES

CHAPTER 3

GLOSSARY OF TERMS

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

sgRNA: Single guide RNA (crRNA-tracrRNA fusion)

PAM: Protospacer adjacent motif

DSB: Double-strand breaks

NHEJ: Non-homologous end joining

HDR: Homology directed repair

ZFNs: Zinc finger nucleases

TALENs: Transcription activator-like effector nucleases

crRNA: CRISPR RNA

tracrRNA: Trans-activating crRNAs

iPSCs: Induced pluripotent stem cells

GATA1: GATA-binding protein-1

ESE: Exonic splicing enhancer

ESS: Exonic splicing silencer

ISE: Intronic splicing enhancer

ISS: Intronic splicing silencer

ssODN: Single-stranded oligonucleotide donor

Indel: Insertions and deletions

ICE: Inference of CRISPR Edits (Hsiau et al., 2018)

TIDE: Tracking of Indels by DEcomposition (Brinkman et al., 2014)

CORRECT: Scarless genome editing (Kwart et al., 2017)

STATEMENT OF RESEARCH

This research was performed in the laboratory of Dr. Vijay G. Sankaran in the Department of Hematology/Oncology at Boston Children's Hospital and Dana-Farber/Boston Children's Cancer and Blood Disorders Center between June 2015 and January 2016. I worked full time from June to August and part time from September to January. I enrolled in the courses HT199.0 Research in Health Sciences and Technology in the fall of 2015 and HT299.0 Research in Health Sciences and Technology II in January 2018. With the exception of Chapter 3, I performed all experiments and generated all figures independently.

ACKNOWLEDGMENTS

I am deeply grateful to Dr. Vijay Sankaran for his extraordinary mentorship. I would like to thank him for his unwavering enthusiasm, dedication, and encouragement throughout my time in research and beyond. I am continuously amazed by how much he cares about my work and how generously he gives his time. He trusted me to explore on my own but was also there every time I needed guidance, whether he was busy on service in the hospital or at home late in the evening. I feel incredibly fortunate to have had the opportunity to study this fascinating blood disorder in his laboratory. I learned so much about how to investigate the mechanisms of disease and writing this thesis has reminded me how much I love science after being away in clinical training for years.

I would like to thank my colleagues in the laboratory for helping me whenever I needed. I am especially grateful to Nour Abdulhay for taking the lead in completing this project and Satish Nandakumar for his suggestions on data analysis.

I would like to thank my classmates Kathy Wang, Paul Dannenberg, and Ken Chang for brainstorming data organization with me. I imagined creating a map of all of my CRISPR editing events by genomic position and Kathy taught me how to do so using a simple spreadsheet! Their friendship and shared enthusiasm for research have been a source of immeasurable joy throughout medical school.

Finally, I would like to thank Harvard-MIT Health Sciences and Technology (HST) and the David G. Nathan Fellowship from Dana-Farber for supporting my research. I am very grateful to Dr. Richard Mitchell and Patty Cunningham for being caring and supportive advisers and Dr. H. Franklin Bunn for introducing me to hematology during the first year of medical school and inspiring me to go into the field.

CHAPTER 1. INTRODUCTION

1.1 Next-generation sequencing heralds an explosion of data

Since the discovery of the structure of DNA, technological advances in sequencing have opened greater and greater territory in the genome for the study of human health and disease. Early studies characterized single genes associated with highly penetrant diseases such as cystic fibrosis, Huntington's disease, and certain types of cancer. Many innovations and collaborations then supported the Human Genome Project, the completion of which in 2003 engendered more complex biological questions and revealed a need for more accessible and robust sequencing methods. Since the mid-2000s, high-throughput sequencing platforms have decreased the cost of human genome sequencing by 50,000-fold and increased capacity by a factor 100-1,000 (Goodwin et al., 2016). These breakthroughs ushered an era of next-generation sequencing and shifted the paradigm to address biological questions on a genome-wide scale.

Next-generation sequencing decodes genomes through a variety of creative approaches. Most widely used are short-read approaches, in which DNA fragments ligated to adapters are clonally amplified on a solid surface and their compositions are detected via a secondary signal, such a fluorophore conjugated to a single nucleotide or a degenerate oligonucleotide or a change in pH associated incorporation of a defined base (Ju et al., 2006; Rothberg et al., 2011). A platform can simultaneously collect information from millions of reaction centers, each with its own clonal DNA template, allowing millions of DNA molecules to be sequenced in parallel. While short-read approaches maximize the number of bases sequenced in the least amount of time, long-read approaches aim to identify longer pieces of contiguous DNA to resolve structurally

complex regions, such as highly repetitive elements and copy number alterations. Instead of clonal amplification, a single DNA molecule progresses through a polymerase fixed to the transparent bottom of a picoliter well and its composition is detected through incorporation of fluorophore-bound nucleotides (Levene et al., 2003; Eid et al., 2009). Alternatively, a system of barcoding can associate fragments sequenced from short-read approaches to create synthetic long reads (McCoy et al., 2014). While whole genome sequencing provides the most comprehensive view, whole exome and targeted sequencing can save time and cost by limiting the genomic material to all coding exons or a custom panel of targeted regions.

The advent of next-generation sequencing has enabled interrogation of nearly every base in the genome. Unlike genome-wide association studies (GWAS), which examine common variants on microarrays, sequencing enables discovery of rare variants with large effects that contribute not only to rare inherited disorders but also common complex diseases not explained by common variants with low to moderate effects (Schork et al., 2009). For example, exome sequencing of two family members with combined hypolipidemia, a rare inherited disorder, identified two distinct nonsense mutations in *ANGPTL3*, which encodes angiopoietin-like 3 protein reported to inhibit lipoprotein lipase and endothelial lipase (Musunuru et al., 2010). Exome sequencing of an Italian family with autosomal dominant amyotrophic lateral sclerosis, another rare inherited disorder, showed a missense mutation in *VCP*, which encodes valosin-containing protein essential for maturation of ubiquitin-containing autophagosomes (Johnson et al., 2010). As for common complex diseases, exome sequencing of a three-generation family with multiple individuals affected by pulmonary arterial hypertension

revealed a frameshift mutation in *CAV1*, which encodes caveolin-1 abundant in the endothelium and other cells of the lung (Austin et al., 2012). Exome sequencing of multiplex families with schizophrenia identified rare coding variants in several genes associated with the N-methyl-D-aspartate (NMDA) receptor, implicating glutamate signaling in the pathogenesis of schizophrenia (Timms et al., 2013). These studies are only a few examples of how rare mutations present fertile ground for further investigation of both Mendelian and complex diseases.

Fueled by next-generation sequencing, the number of inherited phenotypes for which the genetic basis is known has nearly doubled from 2007 to 2013 according to the Online Mendelian Inheritance in Man (OMIM) database (Koboldt et al., 2013). Likewise, the number of known variants in the human genome has risen dramatically over the past decade as seen in Single Nucleotide Polymorphism database (dbSNP). A comparison of dbSNP 135 from October 2011 with dbSNP 137 from June 2012 reveals that most of the recent growth came from variants that are rare (global minor allele frequency < 0.01) or extremely rare (global minor allele frequency < 0.001) in human populations. For many disorders, discovery of the genetic basis has clearly outpaced an understanding of the molecular mechanisms of disease. As sequencing technologies continue to improve and new data continue to accumulate, elucidating the precise relationship between genotype and phenotype and generating functionally and clinically relevant knowledge have become the primary challenge.

1.2 Genome editing holds promise in human disease modeling

Classical approaches to establish gene function include homologous recombination for targeted gene inactivation and RNA inference for targeted gene expression knockdown (Smithies et al., 1985; Thomas and Capecci 1987; McManus and Sharp 2002). Though powerful, both methods have important drawbacks. Homologous recombination is difficult because engineered constructs insert into chromosomal target sites at low efficiency and selection of correctly modified gene is time-consuming and labor-intensive. The newer RNA inference technology is a more efficient alternative to homologous recombination but gene expression knockdown is often transient and incomplete with possible confounding by off-target effects (Qiu et al., 2015).

During the last decade, genome editing, also known as genome engineering, has emerged as a new approach to study genetic alterations. Methods are based on the use of programmable nucleases to induce double-strand breaks (DSBs) at target DNA sites, which stimulate cellular repair mechanisms such as the efficient but error-prone non-homologous end joining (NHEJ) pathway and the high-fidelity but less efficient homology directed repair (HDR) pathway. NHEJ often results in insertions or deletions (indels) at the target site and can knockout gene function via frameshift mutations (**Figure 1B**). HDR is more precise and can incorporate specific mutations, insertions, or deletions via a donor template bearing the desired change and locus-specific homology arms (**Figure 1C**). Genome editing enables manipulation of virtually any sequence in the genome, from single-nucleotide changes to larger insertions and deletions.

The repertoire of tools to induce site-specific DSBs includes zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and clustered

regulatory interspaced short palindromic repeat (CRISPR)/Cas-based RNA-guided DNA endonucleases (Gaj et al., 2013). ZFNs consist of a custom array of site-specific DNA binding domains adapted from zinc finger transcription factors, each of which binds 3-4 bp, fused to the nuclease domain of the bacterial *Fok*I restriction enzyme. Similarly, TALENs consist of a tandem array of DNA binding domains adapted from TAL repeats, each of which binds 1 bp, fused to the nuclease domain of *Fok*I. Unlike ZFNs and TALENs, CRISPR/Cas systems use a combination of guide RNA and Cas nuclease to detect and cleave the target sequence (**Figure 1A**). In bacteria, CRISPR provides acquired immunity by accumulating short segments of foreign DNA, termed "protospacers," from plasmids and phage genomes (Wiedenheft et al., 2012). These sequences are transcribed and processed into short CRISPR RNAs (crRNAs) that recognize specific target DNA and anneal to trans-activating crRNAs (tracrRNAs) that bind Cas nuclease to direct target cleavage. In 2013, four groups demonstrated that expression of *Streptococcus pyogenes* Cas9 and a single guide RNA (sgRNA) fusing crRNA-tracrRNA in mammalian cells induced targeted DSBs (Cong et al., 2013; Mali et al., 2013; Jinek et al., 2013; Cho et al., 2013). The target sites in the mammalian genomes contain a 20 bp sequence that match the protospacer sequence of the guide RNA and an adjacent NGG, or protospacer adjacent motif (PAM) recognized and required by Cas9. The main advantage of CRISPR, compared to ZFNs and TALENs, is the ease with which sgRNAs can be synthesized for targeted modification of different sequences.

Genome editing can replicate or correct mutations in their endogenous context and holds enormous promise in disease modeling. Barth syndrome, for example, is an X-linked cardiac and skeletal mitochondrial myopathy caused by mutations in *TAZ*, which

encodes a protein called tafazzin responsible for normal acylation of cardiolipin, a major phospholipid of the mitochondrial inner membrane. Wang et al. (2014) generated patient-derived cardiomyocytes via reprogramming of skin fibroblasts into induced pluripotent stem cells (iPSCs) and found that the cells assembled sparse and irregular sarcomeres and engineered tissues contracted weakly. They mutated *TAZ* in control iPSCs using CRISPR/Cas9, induced cardiomyocyte differentiation, and demonstrated, elegantly, that the mutations were necessary and sufficient to replicate the phenotypes observed in patient-derived cardiomyocytes. In another study, Pashos et al. (2017) discovered DNA variants in genomic loci associated with regulation of blood lipid traits such as total cholesterol, LDL, HDL, and cholesterol and recapitulated those SNPs via CRISPR/Cas9 in human pluripotent stem cells for functional validation. Other studies demonstrated correction of disease-causing genes using CRISPR/Cas9, such as *CFTR* in intestinal stem cell organoids derived from cystic fibrosis patients (Schwank et al., 2013), *DMD* in iPSCs derived from patients with Duchenne muscular dystrophy (Li et al., 2014), and *HBB* in iPSCs derived from patients with β-thalassemia (Xie et al., 2014). These studies show that genome editing is a powerful approach to model disease and has potential to provide great insight into pathophysiology.

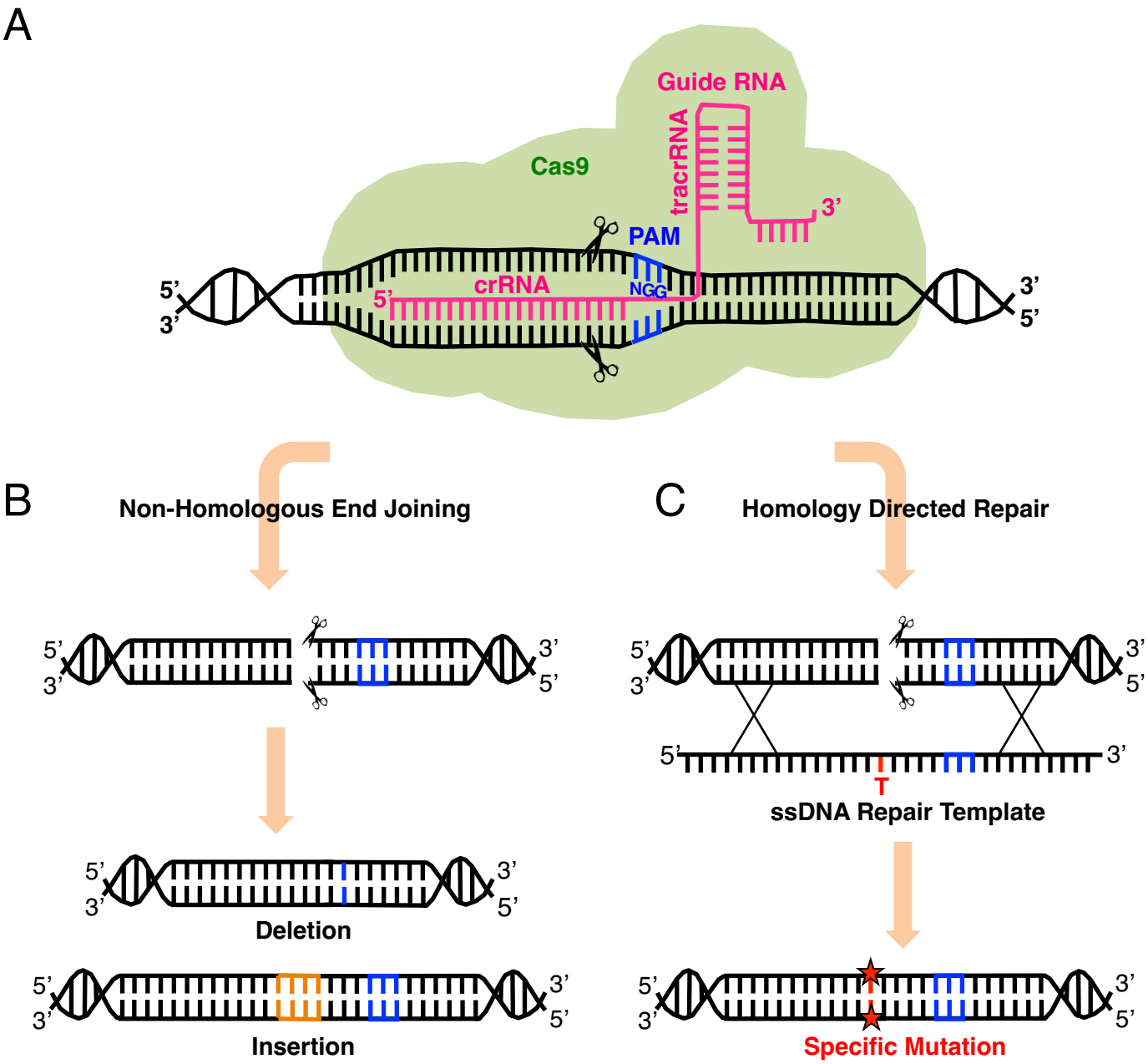


Figure 1. RNA-guided CRISPR-Cas9 enables targeted genome editing
(A) Guide RNA is a short synthetic RNA composed of a scaffold sequence (tracrRNA) that binds to Cas9 nuclease and a 20 nucleotide spacer (crRNA) that provides genomic targeting specificity. The seed sequence refers to 8-10 bases at the 3' end of the spacer, where mismatches are not well tolerated. The ideal targeting sequence is unique in the genome and adjacent to a protospacer adjacent motif (PAM), 5'-NGG-3' when *S. pyogenes* Cas9 is used. Once Cas9 binds to a single guide RNA (sgRNA), it undergoes a conformational change that permits DNA binding. It undergoes a second conformation change upon binding of the Cas9-sgRNA ribonucleoprotein to target DNA that positions the nuclease domains for DNA cleavage at ~3-4 nucleotides upstream of PAM. **(B)** The resulting double-strand breaks (DSB) can be repaired by the efficient but error-prone non-homologous end joining (NHEJ) pathway. This mechanism often results in small nucleotide insertions or deletions (indels) at the DSB site. **(C)** Alternatively, the resulting DSB can be repaired by the high-fidelity but less efficient homology directed repair (HDR) pathway. HDR can be used to generate specific nucleotide changes using a DNA repair template containing the desired mutation. Single-stranded donor oligonucleotides are commonly used for short modifications.

1.3 Patients with an unusual anemia share an intronic mutation in *GATA1*

Two unrelated boys in Germany and the United Kingdom presented with a peculiar form of dyserythropoietic anemia involving persistence of fetal hemoglobin. The patients required transfusions in early infancy but subsequently evolved a milder anemia that worsened in the setting of infections. Common anemias fall into the three broad categories of red cell underproduction, red cell destruction, and excessive blood loss. While some clinical features of the patients indicate hemolysis, others such as dysplastic erythropoiesis on bone marrow biopsy, persistent macrocytosis in the absence of nutritional deficiencies (MCV ~100 fl), elevated fetal hemoglobin post-infancy (~20-23%), and thrombocytopenia with defective platelet function (decreased expression of P-selectin/CD62 and LAMP-3/CD63) suggest atypical hematopoiesis of an intrinsic nature. Extensive workup excluded membranopathies, hemoglobinopathies, and enzymopathies. Curiously, whole exome sequencing was unremarkable but an expanded search including all rare variants in known red blood cell disorder genes revealed a hemizygous point mutation in the fifth intron of *GATA1* (chrX: 48,652,176 C>T in hg19) (**Figure 2**). This gene encodes the hematopoietic master regulatory transcription factor GATA-binding protein-1 (GATA1). The mutation replaces cytosine with thymine 24 nucleotides upstream of the canonical splice acceptor site at 3' end of the intron. The mothers of the two boys are carriers for this mutation on the X chromosome and population studies confirmed its absence from thousands of healthy individuals.

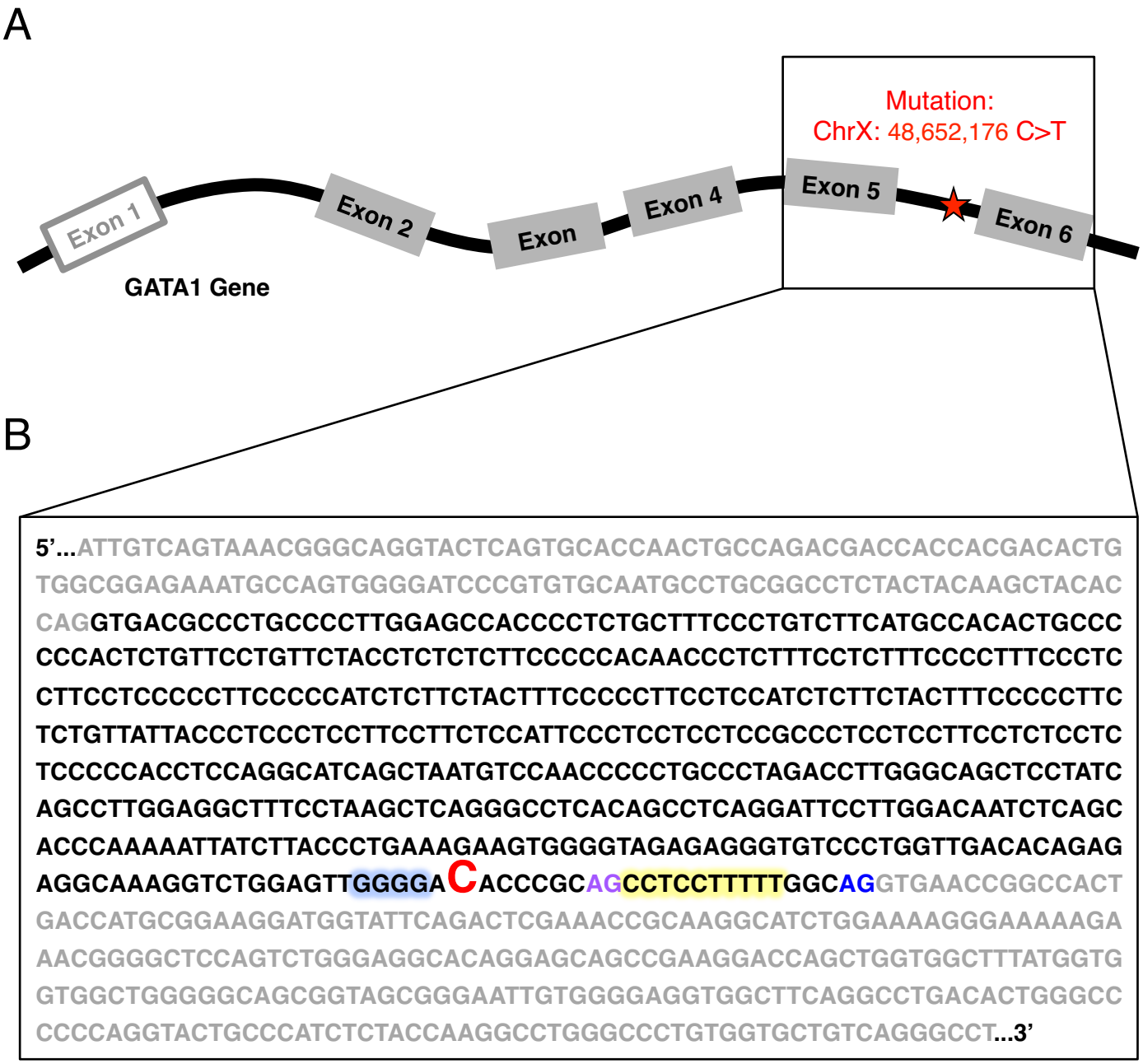


Figure 2. Two unrelated patients with a distinct form of dyserythroplastic anemia have a single nucleotide mutation in the last intron of *GATA1*

(A) The human *GATA1* gene is located on chromosome X and spans 7751 bp. In healthy individuals, alternative splicing produces an mRNA containing exons 1-6 that specifies the full-length GATA1 and a second mRNA with skipping of exon 2 that specifies the truncated GATA1s lacking the N-terminal transactivation domain. Exon 1 (white box) is untranslated and consists of IT and IE components that regulate expression in Sertoli cells and hematopoietic cells, respectively. Two unrelated patients with a distinction presentation of dyserythropietic anemia were noted to have a unique mutation in intron 5 of *GATA1* (ChrX:48,652,176 C>T) after whole exome sequencing was found unrevealing. **(B)** The expansion shows the sequence of exon 5 (gray), intron 5 (black), and exon 6 (gray). The identified mutation (large red C) is located 24 nucleotides upstream of the canonical splice acceptor site (blue). Blue "AG", canonical splice acceptor site. Purple "AG," potential alternative splice acceptor site. Blue highlight, potential intron splicing enhancer. Yellow highlight, potential polyprimidine tract bound by U2-auxiliary factors.

1.4 GATA1 plays an important role in development and disease

Since the first description of GATA1 in 1988, as a protein bound to the 3' enhancer of the human β-globin gene, and its landmark cloning in 1989, studies have illuminated many aspects of its critical function in hematopoiesis (Wall et al., 1988; Tsai et al., 1989). GATA1 plays an essential role in erythroid development. Mouse embryos lacking the transcription factor die from severe anemia between embryonic day 10.5 and 11.5 (Fujiwara et al., 1996). Embryonic stem cells deficient in GATA1 contribute to all tissues in chimeric mice but not red blood cells (Pevny et al., 1991). Erythroid maturation arrests at the proerythroblast stage and precursors die by apoptosis (Pevny et al., 1995; Weiss and Orkin, 1995). In accordance with these observations, GATA1 promotes the survival of erythroid progenitors by activating genes that encode the erythropoietin receptor and the anti-apoptotic protein BCL-X_L (Zon et al., 1991; Gregory et al., 1999). It activates genes important for red cell function, such as those encoding α- and β-globins and heme biosynthesis enzymes (Evans et al., 1988; Orkin, 1992). Concurrently, it represses genes associated with immature, proliferative state, including several core regulators of cell cycle progression (Rylski et al., 2003). In addition, GATA1 is expressed in megakaryocytes, eosinophils, mast cells, and Sertoli cells of the testis and play essential roles in maturation of megakaryocytes and eosinophils (Ferreira et al., 2005) (**Figure 3**). Much of our knowledge about GATA1 is derived from mouse models, grounded in the nearly identical morphology of human and mouse erythroid differentiation. Recent studies, however, note a global divergence in interspecies expression profiles and transcription factor occupancy (Ulirsch et al., 2014). Conserved regulatory regions correspond to a minority of genes essential for the erythroid cell state

(e.g. β-globin, heme biosynthesis enzymes, red cell membrane and surface proteins), as they are likely under strong evolutionary constraint.

The structure of GATA1 corresponds elegantly to its function. The protein has a C-terminal zinc finger that binds to the GATA consensus sequence (A/T)GATA(A/G), an N-terminal zinc finger that stabilizes DNA interactions, and an N-terminal activation domain necessary for transcriptional activation activity at least in exogenous assays (Ferreira et al., 2005). GATA1 can self-associate and interact with a variety of proteins, including cofactors, transcription factors, chromatin-remodeling complexes, and cell cycle regulators. Of particular significance is Friend of GATA1 (FOG1), which binds to the N-terminal zinc finger of GATA1 and synergizes with the transcription factor in gene activation and repression (Fox et al., 1998; Tsang et al., 1997). FOG1 is critical to GATA1 function as mutations that disrupt their interaction block erythropoiesis and compensatory mutations that restore their interaction rescue the phenotype (Crispino et al., 1999). TAL1 is another important cofactor that interacts via LMO2 with the N-terminal zinc finger of GATA1 (Wadman et al., 1997). Unlike FOG1, however, TAL1 and LMO2 cooperate with GATA1 in gene activation but not repression (Tripic et al., 2009; Campbell et al., 2013).

The vast majority of disease-causing mutations perturb the N-terminal domains of GATA1. Missense mutations in the N-terminal zinc finger produce a spectrum of red blood cell and platelet disorders. Not surprisingly, those that severely diminish FOG1-GATA1 interactions (V205M, G208R, D218Y) cause profound anemia and thrombocytopenia and early mortality, whereas those that slightly diminish FOG1-GATA1 interactions (G208S, D218G) or disrupt GATA1-LMO2 interactions (R216Q,

R216W) result in moderate disease (Campbell et al., 2013). Healthy individuals express both full-length GATA1 as well as a truncated form (GATA1s) lacking the N-terminal transactivation domain (Wechsler et al., 2002). Exclusive expression of GATA1s due to loss of the first methionine, splicing errors, and premature termination codon underlies transient abnormal myelopoiesis and myeloid leukemia in children with Down syndrome (Kanezaki et al., 2010). Furthermore, rare mutations that altered the exon 2 donor splice site and the first translation initiation codon of GATA1 led to the discovery that reduced translation of the full-length protein serves as the link between ribosomal protein haploinsufficiency in Diamond-Blackfan anemia and selective aplasia of the erythroid lineage (Sankaran et al., 2012; Ludwig et al., 2014). With a paucity of mutations affecting the C-terminal domain of GATA1, likely reflecting their lethality, the mutation in the last intron may open new understanding of the developmental control and function of this hematopoietic master regulator.

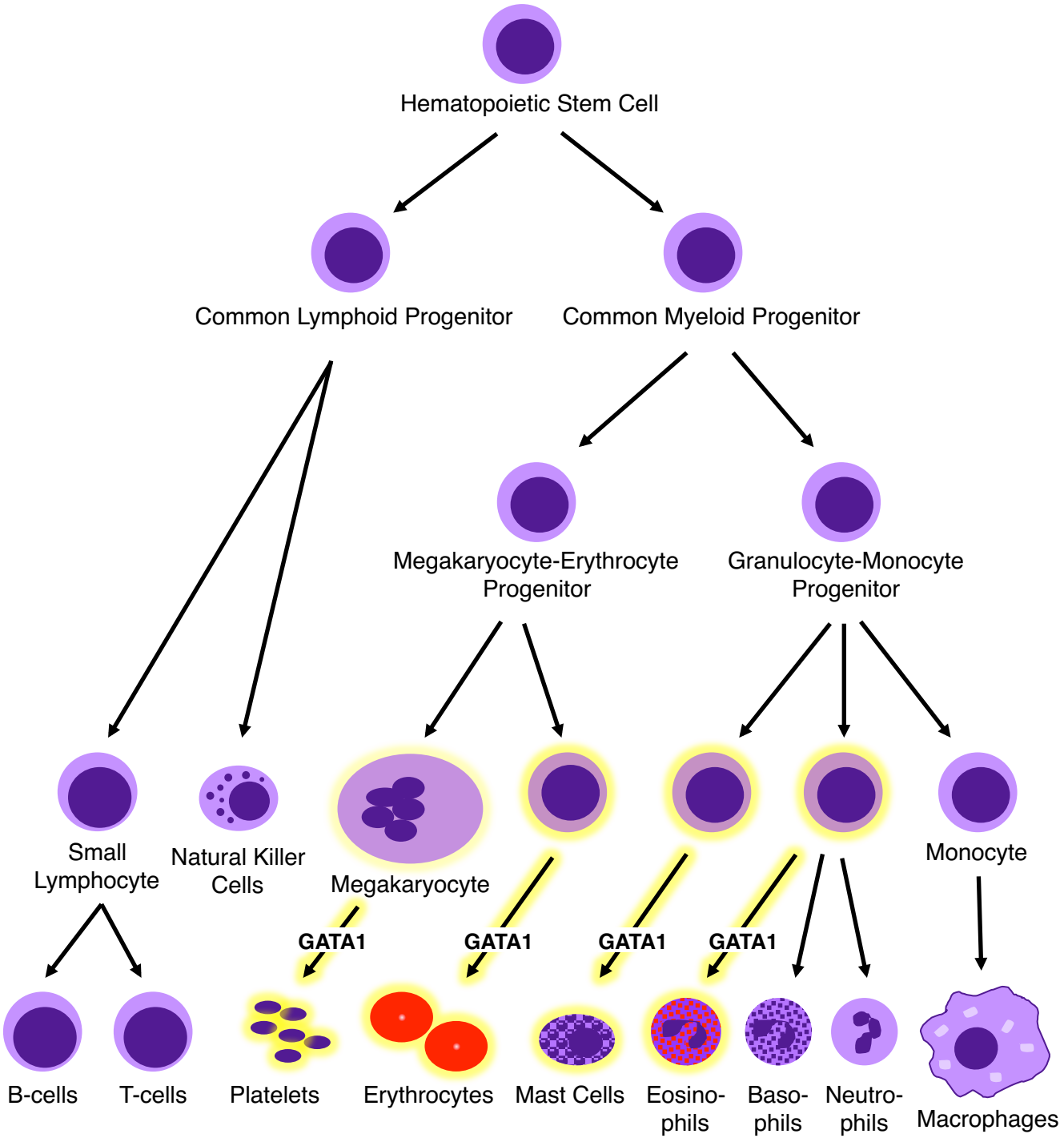


Figure 3. The transcription factor GATA1 plays an essential role in hematopoiesis
Hematopoietic stem cells in the bone marrow can self-renew to maintain their multipotency or give rise to to early progenitors restricted to myeloid or lymphoid differentiation. Common myeloid progenitors mature into erythrocytes, megakaryocytes, granulocytes, and monocytes. Common lymphoid progenitors mature into B-cells, T-cells, and natural killer cells of the immune system. GATA1 plays an essential role in erythroid development. It activates genes important for red cell function, such as those encoding α- and β-globins and heme biosynthesis enzymes, and represses genes associated with immature, proliferative state, including several core regulators of cell cycle progression. In addition to erythroid cells, GATA1 is expressed in megakaryocytes, eosinophils, mast cells, and Sertoli cells of the testis and plays essential roles in the maturation of megakaryocytes and eosinophils.

1.5 Pre-mRNA splicing is a highly regulated process

The mechanism of the intronic mutation is enigmatic and likely involves the modulation of splicing. The distribution of non-coding introns among coding exons was first recognized via the expression of adenovirus (Berget et al., 1997; Chow et al., 1977). Now, it is known that human genes express pre-mRNAs containing eight exons on average (Wang and Cooper, 2007). While exons are typically 50-250 base pairs in length, introns are typically hundreds to thousands of base pairs and account for >90% of the transcription unit on average (Lander et al., 2001; Wang and Burge, 2008). Most human genes undergo alternative splicing or joining of exons that controls gene expression and generates proteomic diversity. *GATA1*, for example, contains two alternative untranslated first exons, IT and IE, and five translated exons, II to VI (Ito et al., 1993; Tsai et al., 1991). In humans, alternative splicing produces an mRNA containing exons I/II/III/IV/V/VI that specifies the full-length GATA1 and a second mRNA containing exons I/III/IV/V/VI that specifies the truncated GATA1s (Halsey et al., 2010). The untranslated exons IT and IE regulate expression in Sertoli cells and hematopoietic cells, respectively (Ito et al., 1993). The long and short forms show similar DNA binding but different transactivation potential, as exon II encodes the first 83 amino acids for the N-terminal transactivation domain (Martin and Orkin, 1990). Even more dramatic is the alternative splicing of *BCL-X*, which produces a smaller mRNA for the pro-apoptotic protein BCL-X_S and a larger mRNA for the anti-apoptotic protein BCL-X_L, the latter of which increases in late erythroid maturation (Motoyama et al., 1995). Splicing is tightly regulated in a developmental stage- and tissue-specific manner. Global analysis during erythroid differentiation show significant changes in alternative splicing, affecting a wide

range of functions from transmembrane receptor activity and chromatin modification to RNA processing and DNA packaging (Cheng et al., 2014). As an error of even a single nucleotide can disrupt the open reading frame and abolish the function of a protein, splicing must occur with extraordinary precision.

Core splicing signals are present in every intron and play instructive roles in splicing reactions. They include the 5' and 3' splice sites at the intron-exon junctions and the branch point sequence approximately 21-34 nucleotides upstream of the 3' end of an intron (Wang and Burge, 2008; Gao et al., 2008). The majority of 5' and 3' splice sites contain the consensus sequences GURAGU and YAG, respectively, where R is a purine (A or G) and Y is a pyrimidine (C or U) (Singh and Cooper, 2012). The first and last two nucleotides of an intron, GU and AG, are almost invariant, whereas other nucleotides are more variable (Wang and Burge, 2008). The human branch point sequence contains a well-conserved adenosine and uridine but other positions are highly degenerative (Gao et al., 2008). Pre-mRNA splicing occurs via two sequential transesterification reactions, where the 2'-hydroxyl group of the branch point adenosine performs a nucleophilic attack on the first nucleotide of the 5' splice site and then the 3'-hydroxyl group of the upstream exon performs a second nucleophilic attack on the last nucleotide at the 3' splice site, joining the exons and releasing the intron lariat (Fica et al., 2013). The spliceosome mediates splice site recognition and catalysis of the reactions. It is composed five small nuclear ribonucleoproteins (snRNPs), U1, U2, U4, U5, and U6 in the major assembly, and approximately 150 proteins (Singh and Cooper, 2012). Initially, U1 binds to the 5' splice site, U2 binds to the branch point sequence, and U2-auxiliary factors bind to the 3' splice site and upstream polypyrimidine tract. U4, U5, and U6 then join the assembly to

remodel the complex. U1 and U4 subsequently exit and U6 catalyzes the transesterification reactions (Fica et al., 2013). The stepwise assembly of snRNPs and interaction of numerous proteins contribute to the precision of splicing.

Nevertheless, core splicing signals contain only about half of the information required for accurate splice site recognition (Lim and Burge, 2001). Many *cis*-regulatory elements are thought to reside in exons and introns and function as splicing enhancers and silencers (Matlin et al., 2005; Wang and Burge, 2008). By recruiting *trans*-acting factors, they promote or inhibit splice site recognition or spliceosome assembly in an additive manner. Exonic splicing enhancers (ESEs) and silencers (ESSs) regulate inclusion of the exon in which they occupy. Most ESEs recruit SR proteins that facilitate spliceosome assembly (Graveley and Maniatis, 1998). ESSs often recruit heterogeneous nuclear ribonucleoproteins (hnRNPs) that can displace snRNP binding, block interactions between snRNPs, or loop out exons (Zhu et al., 2001; Nasim et al., 2002; Sharma et al., 2005). Intronic splicing enhancers (ISEs) and silencers (ISSs) regulate inclusion of adjacent exons. They recruit a variety of *trans*-acting factors including hnRNPs and tissue-specific proteins (Hui et al., 2005; Nakahata and Kawamoto, 2005). While exonic elements are diverse in sequence and large-scale computational and experimental studies have contributed to predictive algorithms, intronic elements are less well characterized and many remain unknown (Chasin, 2007). Reported intronic motifs include G triplet (GGG) and G runs (G_n; $n \geq 3$), which are common in GC-rich introns and function as ISEs, and CA repeats, which can function as ISE or ISS depending on their distance to the upstream exon (McCullough and Berget, 1997; Hui et al., 2005). Importantly, these *cis*-regulatory elements are context-dependent and often have variable effects depending

on their relative positions in the pre-mRNA and gene-specific local secondary structure (Wang and Burge, 2008). Splicing is also intimately linked to the entire process of mRNA transcription, processing, and transport (Hieronymus and Silver, 2004).

1.6 Splicing alterations occur in many diseases

Mutations that disrupt splicing constitute a major cause of disease (Singh and Cooper, 2012). Many act through a single gene by perturbing its splicing code, while some act through multiple genes by perturbing the splicing machinery (Dolatshad et al., 2015). The former are *cis*-acting mutations and include those in the core splicing signals and the regulatory elements that function as splicing enhancers and silencers. The latter are *trans*-acting mutations and include those in the spliceosomal factors and the auxiliary factors that promote or suppress recognition of nearby splice sites.

Cis-acting mutations cause disease through four general mechanisms: 1) constitutive exon skipping or intron retention, 2) activation of cryptic splice sites or pseudoexons, 3) altered inclusion:exclusion ratio of alternative exons, or 4) aberrant splicing mediated by transposable elements (Singh and Cooper, 2012). Mutations in the first category reduce the recognition of 5' or 3' splice site by modifying the consensus sequence of core signals or disrupting the exonic or intronic enhancer elements. Exon skipping is most common but intron retention can occur. These mutations produce an unnatural mRNA that translates into a nonfunctional protein or introduce a premature termination codon (PTC) that targets the mRNA for degradation by nonsense mediated decay (NMD). Examples include mutations in transient abnormal myelopoiesis and myeloid leukemia of Down syndrome that introduce PTCs after the second translation

initiation codon of GATA1 at codon 84, which result in low GATA1s expression by NMD and significantly correlates with disease progression (Kanezaki et al., 2010). Mutations in the second category activate cryptic splice sites, sequences with similar degree of consensus matching as authentic splice sites but rarely if ever used for splicing (Sun and Chasin, 2000). An intronic segment, or pseudoexon, integrates into the mRNA and often leads to disruption of the open reading frame or incorporation of a premature termination codon. This type of error is reported in a variety of diseases, including mild hemophilia A, ocular albinism, Gitelman's syndrome, respiratory disease, and melanoma (Pezeshkpoor et al., 2013; Naruto et al., 2015, Lo et al., 2011; Agrawal et al., 2012; Harland et al., 2001). Mutations in the third category lead to increased or decreased inclusion of an alternative exon by disrupting splicing enhancers or silencers. Alternative splicing is tightly regulated to produce normal ratios of natural mRNA isoforms. Disruption of tau protein isoforms, for example, underlies frontotemporal dementia and Parkinsonism linked to chromosome 17 (D' Souza et al., 1999). Finally, insertion of transposable elements leads to aberrant splicing through alternate use of splice sites and generally produces nonfunctional proteins. In Fukuyama muscular dystrophy, inclusion of a new exon from a transposable element in the *fukutin* gene results in a mislocalized and nonfunctional protein (Taniguchi-Ikeda et al., 2011)

The intronic mutation in GATA1 replaces cytosine with thymine 24 nucleotides upstream of the last intron-exon boundary (**Figure 2B**). It does not disrupt the 5' splice donor site, the 3' splice acceptor site, or the branch point adenosine. It precedes a pyrimidine-rich region that could represent the polypyrimidine tract bound by U2-auxiliary factors but is unlikely to disrupt these interactions, as C is substituted with U,

another pyrimidine. The mutation is 2 nucleotides downstream of a G_4 run, a potential ISE, and 6 nucleotides upstream of a CAG motif, which may represent a potential alternative splice acceptor site. Its mechanism is not immediately revealing and exon skipping, intron retention, and activation of a cryptic acceptor site are all plausible. I hypothesized that genome editing is an effective approach to model this new disease and splicing alteration leads reduced expression of functional GATA1 that may produce the clinical features of the unusual anemia observed in our patients.

REFERENCES

Agrawal, A., Hamvas, A., Cole, F.S., Wambach, J.A., Wegner, D., Coghill, C., Harrison, K., and Nogee, L.M. (2012). An intronic ABCA3 mutation that is responsible for respiratory disease. Pediatr. Res. *71*, 633–637.

Austin, E.D., Ma, L., LeDuc, C., Berman Rosenzweig, E., Borczuk, A., Phillips, J.A., Palomero, T., Sumazin, P., Kim, H.R., Talati, M.H., et al. (2012). Whole exome sequencing to identify a novel gene (caveolin-1) associated with human pulmonary arterial hypertension. Circ Cardiovasc Genet *5*, 336–343.

Berget, S.M., Moore, C., and Sharp, P.A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc. Natl. Acad. Sci. U.S.A. *74*, 3171–3175.

Campbell, A.E., Wilkinson-White, L., Mackay, J.P., Matthews, J.M., and Blobel, G.A. (2013). Analysis of disease-causing GATA1 mutations in murine gene complementation systems. Blood *121*, 5218–5227.

Chasin, L.A. (2007). Searching for splicing motifs. Adv. Exp. Med. Biol. *623*, 85–106.

Cheng, A.W., Shi, J., Wong, P., Luo, K.L., Trepman, P., Wang, E.T., Choi, H., Burge, C.B., and Lodish, H.F. (2014). Muscleblind-like 1 (Mbnl1) regulates pre-mRNA alternative splicing during terminal erythropoiesis. Blood *124*, 598–610.

Cho, S.W., Kim, S., Kim, J.M., and Kim, J.-S. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat. Biotechnol. *31*, 230–232.

Chow, L.T., Gelinas, R.E., Broker, T.R., and Roberts, R.J. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell *12*, 1–8.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Crispino, J.D., Lodish, M.B., MacKay, J.P., and Orkin, S.H. (1999). Use of altered specificity mutants to probe a specific protein-protein interaction in differentiation: the GATA-1:FOG complex. Mol. Cell *3*, 219–228.

Dolatshad, H., Pellagatti, A., Fernandez-Mercado, M., Yip, B.H., Malcovati, L., Attwood, M., Przychodzen, B., Sahgal, N., Kanapin, A.A., Lockstone, H., et al. (2015). Disruption of SF3B1 results in deregulated expression and splicing of key genes and pathways in myelodysplastic syndrome hematopoietic stem and progenitor cells. Leukemia *29*, 1092–1103.

D'Souza, I., Poorkaj, P., Hong, M., Nochlin, D., Lee, V.M., Bird, T.D., and Schellenberg, G.D. (1999). Missense and silent tau gene mutations cause frontotemporal dementia with parkinsonism-chromosome 17 type, by affecting multiple alternative RNA splicing regulatory elements. Proc. Natl. Acad. Sci. U.S.A. *96*, 5598–5603.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138.

Evans, T., Reitman, M., and Felsenfeld, G. (1988). An erythrocyte-specific DNA-binding factor recognizes a regulatory sequence common to all chicken globin genes. Proc. Natl. Acad. Sci. U.S.A. *85*, 5976–5980.

Ferreira, R., Ohneda, K., Yamamoto, M., and Philipsen, S. (2005). GATA1 function, a paradigm for transcription factors in hematopoiesis. Mol. Cell. Biol. *25*, 1215–1227.

Fica, S.M., Tuttle, N., Novak, T., Li, N.-S., Lu, J., Koodathingal, P., Dai, Q., Staley, J.P., and Piccirilli, J.A. (2013). RNA catalyses nuclear pre-mRNA splicing. Nature *503*, 229–234.

Fox, A.H., Kowalski, K., King, G.F., Mackay, J.P., and Crossley, M. (1998). Key residues characteristic of GATA N-fingers are recognized by FOG. J. Biol. Chem. *273*, 33595–33603.

Fujiwara, Y., Browne, C.P., Cunniff, K., Goff, S.C., and Orkin, S.H. (1996). Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. Proc. Natl. Acad. Sci. U.S.A. *93*, 12355–12358.

Gaj, T., Gersbach, C.A., and Barbas, C.F. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends in Biotechnology *31*, 397–405.

Gao, K., Masuda, A., Matsuura, T., and Ohno, K. (2008). Human branch point consensus sequence is yUnAy. Nucleic Acids Res. *36*, 2257–2267.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351.

Graveley, B.R., and Maniatis, T. (1998). Arginine/serine-rich domains of SR proteins can function as activators of pre-mRNA splicing. Mol. Cell *1*, 765–771.

Gregory, T., Yu, C., Ma, A., Orkin, S.H., Blobel, G.A., and Weiss, M.J. (1999). GATA-1 and erythropoietin cooperate to promote erythroid cell survival by regulating bcl-xL expression. Blood *94*, 87–96.

Halsey, C., Tunstall, O., Gibson, B., Roberts, I., and Graham, G. (2010). Role of GATA-1s in early hematopoiesis and differences between alternative splicing in human and murine GATA-1. Blood *115*, 3415–3416.

Harland, M., Mistry, S., Bishop, D.T., and Bishop, J.A. (2001). A deep intronic mutation in CDKN2A is associated with disease in a subset of melanoma pedigrees. Hum. Mol. Genet. *10*, 2679–2686.

Hieronymus, H., and Silver, P.A. (2004). A systems view of mRNP biology. Genes Dev.

18, 2845–2860.

Hui, J., Hung, L.-H., Heiner, M., Schreiner, S., Neumüller, N., Reither, G., Haas, S.A., and Bindereif, A. (2005). Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. EMBO J. *24*, 1988–1998.

Ito, E., Toki, T., Ishihara, H., Ohtani, H., Gu, L., Yokoyama, M., Engel, J.D., and Yamamoto, M. (1993). Erythroid transcription factor GATA-1 is abundantly transcribed in mouse testis. Nature *362*, 466–468.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. Elife *2*, e00471.

Johnson, J.O., Mandrioli, J., Benatar, M., Abramzon, Y., Van Deerlin, V.M., Trojanowski, J.Q., Gibbs, J.R., Brunetti, M., Gronka, S., Wuu, J., et al. (2010). Exome sequencing reveals VCP mutations as a cause of familial ALS. Neuron *68*, 857–864.

Ju, J., Kim, D.H., Bi, L., Meng, Q., Bai, X., Li, Z., Li, X., Marma, M.S., Shi, S., Wu, J., et al. (2006). Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc. Natl. Acad. Sci. U.S.A. *103*, 19635–19640.

Kanezaki, R., Toki, T., Terui, K., Xu, G., Wang, R., Shimada, A., Hama, A., Kanegane, H., Kawakami, K., Endo, M., et al. (2010). Down syndrome and GATA1 mutations in transient abnormal myeloproliferative disorder: mutation classes correlate with progression to myeloid leukemia. Blood *116*, 4631–4638.

Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K., and Mardis, E.R. (2013). The Next-Generation Sequencing Revolution and Its Impact on Genomics. Cell *155*, 27–38.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G., and Webb, W.W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. Science *299*, 682–686.

Li, H.L., Fujimoto, N., Sasakawa, N., Shirai, S., Ohkame, T., Sakuma, T., Tanaka, M., Amano, N., Watanabe, A., Sakurai, H., et al. (2015). Precise Correction of the Dystrophin Gene in Duchenne Muscular Dystrophy Patient Induced Pluripotent Stem Cells by TALEN and CRISPR-Cas9. Stem Cell Reports *4*, 143–154.

Lim, L.P., and Burge, C.B. (2001). A computational analysis of sequence features involved in recognition of short introns. Proc. Natl. Acad. Sci. U.S.A. *98*, 11193–11198.

Lo, Y.-F., Nozu, K., Iijima, K., Morishita, T., Huang, C.-C., Yang, S.-S., Sytwu, H.-K., Fang, Y.-W., Tseng, M.-H., and Lin, S.-H. (2011). Recurrent deep intronic mutations in

the SLC12A3 gene responsible for Gitelman's syndrome. Clin J Am Soc Nephrol *6*, 630–639.

Ludwig, L.S., Gazda, H.T., Eng, J.C., Eichhorn, S.W., Thiru, P., Ghazvinian, R., George, T.I., Gotlib, J.R., Beggs, A.H., Sieff, C.A., et al. (2014). Altered translation of GATA1 in Diamond-Blackfan anemia. Nat. Med. *20*, 748–753.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-Guided Human Genome Engineering via Cas9. Science *339*, 823–826.

Martin, D.I., and Orkin, S.H. (1990). Transcriptional activation and DNA binding by the erythroid factor GF-1/NF-E1/Eryf 1. Genes Dev. *4*, 1886–1898.

Matlin, A.J., Clark, F., and Smith, C.W.J. (2005). Understanding alternative splicing: towards a cellular code. Nat. Rev. Mol. Cell Biol. *6*, 386–398.

McCoy, R.C., Taylor, R.W., Blauwkamp, T.A., Kelley, J.L., Kertesz, M., Pushkarev, D., Petrov, D.A., and Fiston-Lavier, A.-S. (2014). Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PLoS ONE *9*, e106689.

McCullough, A.J., and Berget, S.M. (1997). G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. Mol. Cell. Biol. *17*, 4562–4571.

McManus, M.T., and Sharp, P.A. (2002). Gene silencing in mammals by small interfering RNAs. Nat. Rev. Genet. *3*, 737–747.

Motoyama, N., Wang, F., Roth, K.A., Sawa, H., Nakayama, K., Nakayama, K., Negishi, I., Senju, S., Zhang, Q., and Fujii, S. (1995). Massive cell death of immature hematopoietic cells and neurons in Bcl-x-deficient mice. Science *267*, 1506–1510.

Musunuru, K., Pirruccello, J.P., Do, R., Peloso, G.M., Guiducci, C., Sougnez, C., Garimella, K.V., Fisher, S., Abreu, J., Barry, A.J., et al. (2010). Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. N. Engl. J. Med. *363*, 2220–2227.

Nakahata, S., and Kawamoto, S. (2005). Tissue-dependent isoforms of mammalian Fox-1 homologs are associated with tissue-specific splicing activities. Nucleic Acids Res. *33*, 2078–2089.

Naruto, T., Okamoto, N., Masuda, K., Endo, T., Hatsukawa, Y., Kohmoto, T., and Imoto, I. (2015). Deep intronic GPR143 mutation in a Japanese family with ocular albinism. Sci Rep *5*, 11334.

Nasim, F.-U.H., Hutchison, S., Cordeau, M., and Chabot, B. (2002). High-affinity hnRNP A1 binding sites and duplex-forming inverted repeats have similar effects on 5'

splice site selection in support of a common looping out and repression mechanism. RNA *8*, 1078–1089.

Orkin, S.H. (1992). GATA-binding transcription factors in hematopoietic cells. Blood *80*, 575–581.

Pashos, E.E., Park, Y., Wang, X., Raghavan, A., Yang, W., Abbey, D., Peters, D.T., Arbelaez, J., Hernandez, M., Kuperwasser, N., et al. (2017). Large, Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like Cells Reveal Functional Genetic Variation at Blood Lipid-Associated Loci. Cell Stem Cell *20*, 558–570.e10.

Pevny, L., Simon, M.C., Robertson, E., Klein, W.H., Tsai, S.F., D'Agati, V., Orkin, S.H., and Costantini, F. (1991). Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. Nature *349*, 257–260.

Pevny, L., Lin, C.S., D'Agati, V., Simon, M.C., Orkin, S.H., and Costantini, F. (1995). Development of hematopoietic cells lacking transcription factor GATA-1. Development *121*, 163–172.

Pezeshkpoor, B., Zimmer, N., Marquardt, N., Nanda, I., Haaf, T., Budde, U., Oldenburg, J., and El-Maarri, O. (2013). Deep intronic "mutations" cause hemophilia A: application of next generation sequencing in patients without detectable mutation in F8 cDNA. J. Thromb. Haemost. *11*, 1679–1687.

Qiu, S., Adema, C.M., and Lane, T. (2005). A computational study of off-target effects of RNA interference. Nucleic Acids Res. *33*, 1834–1847.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature *475*, 348–352.

Rylski, M., Welch, J.J., Chen, Y.-Y., Letting, D.L., Diehl, J.A., Chodosh, L.A., Blobel, G.A., and Weiss, M.J. (2003). GATA-1-mediated proliferation arrest during erythroid maturation. Mol. Cell. Biol. *23*, 5031–5042.

Sankaran, V.G., Ghazvinian, R., Do, R., Thiru, P., Vergilio, J.-A., Beggs, A.H., Sieff, C.A., Orkin, S.H., Nathan, D.G., Lander, E.S., et al. (2012). Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. Journal of Clinical Investigation *122*, 2439–2443.

Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. rare allele hypotheses for complex diseases. Curr. Opin. Genet. Dev. *19*, 212–219.

Schwank, G., Koo, B.-K., Sasselli, V., Dekkers, J.F., Heo, I., Demircan, T., Sasaki, N., Boymans, S., Cuppen, E., van der Ent, C.K., et al. (2013). Functional repair of CFTR by CRISPR/Cas9 in intestinal stem cell organoids of cystic fibrosis patients. Cell Stem Cell *13*, 653–658.

Sharma, S., Falick, A.M., and Black, D.L. (2005). Polypyrimidine tract binding protein blocks the 5' splice site-dependent assembly of U2AF and the prespliceosomal E complex. Mol. Cell *19*, 485–496.

Singh, R.K., and Cooper, T.A. (2012). Pre-mRNA splicing in disease and therapeutics. Trends Mol Med *18*, 472–482.

Smithies, O., Gregg, R.G., Boggs, S.S., Koralewski, M.A., and Kucherlapati, R.S. (1985). Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination. Nature *317*, 230–234.

Sun, H., and Chasin, L.A. (2000). Multiple splicing defects in an intronic false exon. Mol. Cell. Biol. *20*, 6414–6425.

Taniguchi-Ikeda, M., Kobayashi, K., Kanagawa, M., Yu, C., Mori, K., Oda, T., Kuga, A., Kurahashi, H., Akman, H.O., DiMauro, S., et al. (2011). Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy. Nature *478*, 127–131.

Thomas, K.R., and Capecchi, M.R. (1987). Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. Cell *51*, 503–512.

Timms, A.E., Dorschner, M.O., Wechsler, J., Choi, K.Y., Kirkwood, R., Girirajan, S., Baker, C., Eichler, E.E., Korvatska, O., Roche, K.W., et al. (2013). Support for the N-methyl-D-aspartate receptor hypofunction hypothesis of schizophrenia from exome sequencing in multiplex families. JAMA Psychiatry *70*, 582–590.

Tripic, T., Deng, W., Cheng, Y., Zhang, Y., Vakoc, C.R., Gregory, G.D., Hardison, R.C., and Blobel, G.A. (2009). SCL and associated proteins distinguish active from repressive GATA transcription factor complexes. Blood *113*, 2191–2201.

Tsai, S.F., Martin, D.I., Zon, L.I., D'Andrea, A.D., Wong, G.G., and Orkin, S.H. (1989). Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. Nature *339*, 446–451.

Tsai, S.F., Strauss, E., and Orkin, S.H. (1991). Functional analysis and in vivo footprinting implicate the erythroid transcription factor GATA-1 as a positive regulator of its own promoter. Genes Dev. *5*, 919–931.

Tsang, A.P., Visvader, J.E., Turner, C.A., Fujiwara, Y., Yu, C., Weiss, M.J., Crossley, M., and Orkin, S.H. (1997). FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. Cell *90*, 109–119.

Wadman, I.A., Osada, H., Grütz, G.G., Agulnick, A.D., Westphal, H., Forster, A., and Rabbitts, T.H. (1997). The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. EMBO J. *16*, 3145–3157.

Wall, L., deBoer, E., and Grosveld, F. (1988). The human beta-globin gene 3' enhancer contains multiple binding sites for an erythroid-specific protein. Genes Dev. *2*, 1089–1100.

Wang, G.-S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat. Rev. Genet. *8*, 749–761.

Wang, Z., and Burge, C.B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA *14*, 802–813.

Wang, G., McCain, M.L., Yang, L., He, A., Pasqualini, F.S., Agarwal, A., Yuan, H., Jiang, D., Zhang, D., Zangi, L., et al. (2014). Modeling the mitochondrial cardiomyopathy of Barth syndrome with induced pluripotent stem cell and heart-on-chip technologies. Nat. Med. *20*, 616–623.

Wechsler, J., Greene, M., McDevitt, M.A., Anastasi, J., Karp, J.E., Le Beau, M.M., and Crispino, J.D. (2002). Acquired mutations in GATA1 in the megakaryoblastic leukemia of Down syndrome. Nat. Genet. *32*, 148–152.

Weiss, M.J., and Orkin, S.H. (1995). Transcription factor GATA-1 permits survival and maturation of erythroid precursors by preventing apoptosis. Proc. Natl. Acad. Sci. U.S.A. *92*, 9623–9627.

Wiedenheft, B., Sternberg, S.H., and Doudna, J.A. (2012). RNA-guided genetic silencing systems in bacteria and archaea. Nature *482*, 331–338.

Xie, F., Ye, L., Chang, J.C., Beyer, A.I., Wang, J., Muench, M.O., and Kan, Y.W. (2014). Seamless gene correction of β-thalassemia mutations in patient-specific iPSCs using CRISPR/Cas9 and *piggyBac*. Genome Research *24*, 1526–1533.

Zhu, J., Mayeda, A., and Krainer, A.R. (2001). Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. Mol. Cell *8*, 1351–1361.

Zon, L.I., Youssoufian, H., Mather, C., Lodish, H.F., and Orkin, S.H. (1991). Activation of the erythropoietin receptor promoter by transcription factor GATA-1. Proc. Natl. Acad. Sci. U.S.A. *88*, 10638–10641.

CHAPTER 2. CRISPR-Cas9 Recapitulates the *GATA1* Intronic Mutation of the Dyserythropoietic Anemia with Additional Modifications

RESULTS & DISCUSSION

2.1 CRISPR-Cas9 offers the possibility of creating an isogenic human erythroid cell line with the mutation of interest in *GATA1*

To model the novel dyserythropoietic anemia in a tractable system, I aimed to create human erythroid cell lines with the single nucleotide change in intron 5 of *GATA1* identified on targeted mutation analysis (ChrX:48,652,176 C>T in hg19) (**Figure 2**). The K562 cell line derived from a patient with chronic myelogenous leukemia in blast crisis has been used in numerous applications, including the differentiation of the hematopoietic cells (Lozzio and Lozzio, 1975; Naumann et al., 2000). Cellular models of human origin are preferable to mouse models as our lab has previously demonstrated interspecies transcriptional and splicing divergence in erythropoiesis (Ulirsch et al., 2014). Notably, the occupancy sites of GATA1 and many other transcription factors are significantly more conserved between primary cells and erythroid cell lines derived from the same species than across species.

My method of choice was to perform targeted genome editing via the RNA-guided CRISPR-Cas9 nuclease system. This system can be leveraged to generate the single nucleotide mutation in *GATA1* in the genome of K562 cells by simultaneous delivery of a Cas9 nuclease, a single-guide RNA (sgRNA), and an exogenous DNA repair template containing thymine in place of cytosine at the desired position. Once expressed in cells, Cas9 may induce DSBs near the intended site of mutation to enable incorporation of the single nucleotide change via HDR (**Figure 1C**). CRISPR-Cas9

provides an ideal approach to study the molecular mechanisms of this enigmatic mutation as it leaves the rest of the genome unperturbed.

In my initial experiment, I designed sgRNAs 1-4 using the Optimized CRISPR Design tool developed by the Feng Zhang Lab at MIT (**Figure 4A-B**). I prioritized guides that minimized the distance from the Cas9-mediated DSBs to the mutation site and the probability of off-target binding. All four guides have a computed score of 50 and above and are deemed high quality candidates based on faithfulness of on-target activity. The distance from the DSBs to the mutation site ranged from 10 to 27 bp (**Figure 4C**). As all four sgRNAs are oriented in the sense direction, I designed a sense single-stranded oligonucleotide donor (ssODN) as the repair template to avoid Cas9-mediated cleavage (Chen et al., 2011). The ssODN contains the thymine mutation with flanking sequences of 90 bp homologous to the target region (Ran et al., 2013). After cloning the sgRNAs into a plasmid vector, I selected sgRNAs 3 and 4 for direct nucleofection of K526 cells with Cas9 and ssODN (**Figure 6A**). These two guides have the shortest distance from the Cas9 cleavage site to the intended mutation, at 15 and 10 bp, respectively. The plasmid vector for Cas9 carries GFP and puromycin resistance genes, enabling both visual and chemical selection of successfully nucleofected cells (**Figure 6B-C**). After 48 hours of puromycin treatment, FACS analysis showed that 65.1% of an untreated K562 cell sample and 88.9% of a nucleofected K562 cell sample were propidium iodide (PI)-negative, indicating greater survival in the latter group (**Figure 6C**). Many control cells remain PI-negative at this time point, as it may take longer for puromycin to cause cell death. Of the 88.9% of nucleofected K562 cells that are PI-negative, 92.2% were GFP+, indicating that most of the living cells are expressing Cas9. To isolate isogenic cell lines,

I then performed serial dilutions based on the FACS estimates to achieve single cell plating for clonal expansion. To screen for the intended mutation, I purified DNA from 26 clones nucleofected with sgRNA 3 and 20 clones nucleofected with sgRNA 4 for PCR amplification of the *GATA1* target region and sequence analysis. Unexpectedly, all clones showed wild-type sequence of the target region with no evidence of genome editing in the form of insertion, deletion, or substitution (**Figure 7A**). As impaired *GATA1* function may reduce cell survival, I then screened 9 slowest-proliferating clones nucleofected with sgRNA 4, which has the shortest cut-to-mutation distance and likely the highest probability for mutation incorporation (Kwart et al., 2017). Again, all clones showed wild-type sequence of the *GATA1* target region (**Figure 7A**). Though sgRNAs typically work except on rare occasions for reasons not yet known (Ran et al., 2013), these results necessitate functional analysis of my sgRNAs.

2.2 Surveyor nuclease assay identifies one guide RNA for targeted editing of *GATA1*

Surveyor nuclease is a special endonuclease that cleaves DNA with high specificity at sites of base substitution mismatch (Qiu et al. 2004). I harnessed this function to determine the ability of my sgRNAs to create targeted genomic modifications in *GATA1* via CRIPSR-Cas9. For functional analysis of my sgRNAs, 293T cells provide a better system than K562 cells as they have high transfection efficiency and do not require GATA1 for survival. After transfecting 293T cells with sgRNAs 1-4 and GFP-Cas9, I performed puromycin selection and harvested genomic DNA from each population for PCR amplification of the *GATA1* target region. If CRISPR-mediated modifications were present, heat denaturation and reannealing of the amplified

heterogeneous DNA population would lead to formation of DNA heteroduplexes with mismatched bases at the site of Cas9 cleavage due to indels from NHEJ. Surveyor nuclease would then cleave the DNA heteroduplexes at the sites of mismatches to generate products smaller than the original PCR amplicon that can be visualized on a gel.

Using this assay, I found that sgRNAs 1-4 did not lead to CRISPR-mediated modifications in *GATA1*. In each reaction, only one band corresponding to the original PCR amplicon (397 bp) is present, suggesting that there were no indels in the target region and therefore no formation of DNA heteroduplexes or surveyor nuclease cleavage products (**Figure 5C**). The same result is seen when 293T cells are transfected with an sgRNA targeting another gene, *ALAS2*, which encodes delta-aminolevulinate synthase 2 that catalyzes the first step in the heme biosynthesis pathway, or without an sgRNA (Cas9 only). Control G and Control C are plasmids with inserts that differ at a single base pair. As a positive control, I amplified a region containing this single base pair change and hybridized equal amounts of PCR products from both plasmids. As a negative control, I hybridized equal amounts of PCR products from Control C only. As expected, 3 bands were present in the positive control, corresponding to G/G and C/C homoduplexes that were not cleaved by the surveyor nuclease (632 bp) and G/C and C/G heteroduplexes that were cleaved into products of 416 and 217 bp. Only 1 band is present in the negative control, corresponding to C/C homoduplexes that were not cleaved by the surveyor nuclease. These controls validate the function of the surveyor nuclease.

To ensure that the surveyor nuclease assay can confirm genome editing following CRISPR and that sgRNAs 1-4 did not lead to modifications of *GATA1* due to their targeting sequences, I transfected 293T cells with an sgRNA targeting *ALAS2* and

amplified the target region in *ALAS2* for heteroduplex formation and surveyor nuclease digestion (**Figure 5B**). This sgRNA has been validated in prior CRISPR studies in our lab. As expected, the surveyor nuclease assay confirmed targeted editing of *ALAS2* with 3 bands corresponding to the wild-type PCR amplicon (713 bp) and digested products (387 and 326 bp). If only Cas9 were transfected or if the amplified and reannealed products were not treated by the surveyor nuclease, only 1 band corresponding to the original PCR amplicon is present. These results suggest that *GATA1* editing did not occur with sgRNAs 1-4 due to their targeting sequences and not other variables in CRISPR. One possible reason is that sgRNAs 1-3 contain five consecutive thymine bases (**Figure 4A**). A poly-T signal can cause catalytic inactivation and backtracking of RNA Pol III (Nielsen et al., 2013), leading to premature transcription termination of these sgRNAs. sgRNA 4 does not have a poly-T signal but has the most number of estimated off-target sites, which may contribute to the lack of *GATA1* editing.

Next, I designed 5 new sgRNAs targeting *GATA1*. I included 3 sgRNAs oriented in the antisense direction to avoid the poly-T signal (sgRNAs 5-7) and 2 sgRNAs oriented in the sense direction targeting a region further upstream (sgRNAs 8-9). The distance from the Cas9-mediated DSBs to the mutation site ranged from 15 to 34 bp (**Figure 4B**). After cloning these new guides into a plasmid vector, I transfected 293T cells and performed functional analysis with the surveyor nuclease. Only sgRNA 5 appeared to generate CRISPR-mediated modifications in *GATA1*, with evidence of surveyor nuclease digested products at 250 bp and 147 bp (**Figure 6D**). The 147 bp band is nearly undetectable, perhaps due to its small mass. sgRNAs 6-9 appear non-functional, with no clear evidence of surveyor nuclease digested products. Fortunately, among all

guides, sgRNA 5 has the lowest estimated number of off-target sites (**Figure 4A**). However, at 18 bp from the Cas9-mediated cleavage site to the intended mutation site (**Figure 4C**), generating a homozygous mutation may be inefficient. One study estimates that <10% of clones with evidence of HDR incorporated a homozygous mutation at this distance (Kwart et al., 2017).

These experiments indicate that functional validation of sgRNA is a key step in performing genome editing with CRISPR-Cas9. They also highlight several limitations in finding an ideal guide, including 1) availability of unique targeting sequences adjacent to PAM sequences in the genomic region of interest to minimize off-target binding and 2) distance from the Cas9-mediated cut to the mutation site to optimize the likelihood of intended modification. Computational analysis may generate multiple promising sgRNA candidates but functional analysis may narrow the possibilities significantly. Antisense sgRNA targeting the transcriptionally active DNA strand may result in higher Cas9-mediated cleavage and repair activity than sense sgRNA targeting the transcriptionally inactive DNA strand (Song and Stieger 2017).

A

sgRNA	Target Sequence	Score*	Offtargets	Direction
1	5'-CTCCTTTTTGGCAGGTGAACCGG-3'	68	211	Sense
2	5'-CCCGCAGCCTCCTTTTTGGCAGG-3'	68	199	Sense
3	5'-GACACCCGCAGCCTCCTTTTTGG-3'	67	170	Sense
4	5'-GAGAGGCAAAGGTCTGGAGTTGG-3'	50	444	Sense
5	5'-GGCCGGTTCACCTGCCAAAAAGG-3'	82	78	Antisense
6	5'-CGGTTCACCTGCCAAAAAGGAGG-3'	70	105	Antisense
7	5'-TTTGCCTCTCTGTGTCAACCAGG-3'	68	325	Antisense
8	5'-TGTCCCTGGTTGACACAGAGAGG-3'	52	309	Sense
9	5'-GACACAGAGAGGCAAAGGTCTGG-3'	50	478	Sense

*Inverse likelihood of off-target binding

B

5'...CAGCACCCAAAAATTATCTTACCCTGAAAGAAGTGGGGTAGAGAGGGTGTCCCTGGTTGA
CACAGAGAGGCAAAGGTCTGGAGTTGGGGACACCCGCAGCCTCCTTTTTGGCAGGTGAACC
GGCCACTGACCATGCGGAAGGATGGTATTCAGACTCGAAACCGCAAGGCATCTGGAAAAG...3'

C

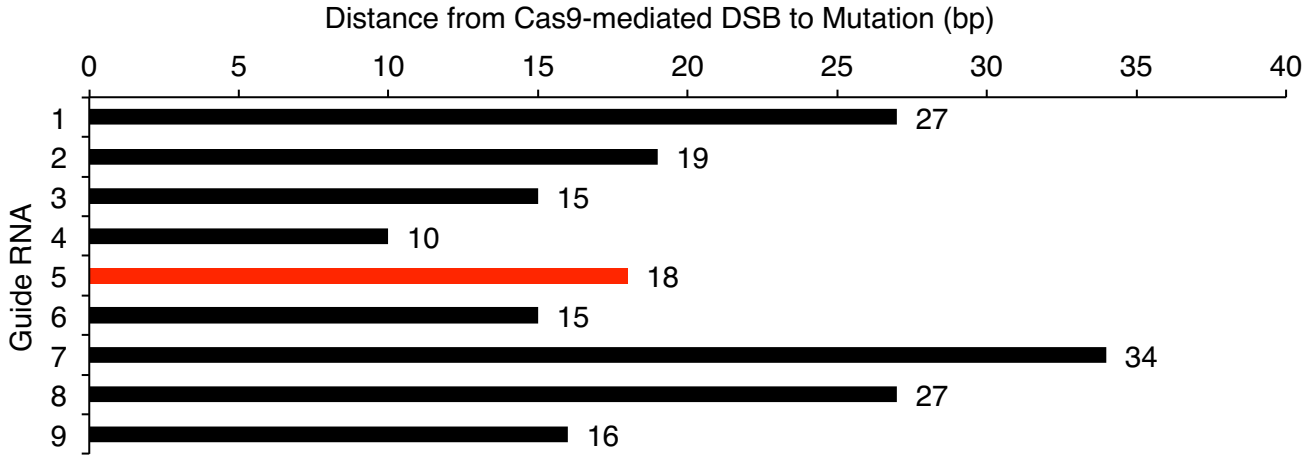


Figure 4. In silico analysis identifies multiple guide RNAs for targeted editing of *GATA1*
(A) Optimized CRISPR Design from Feng Zhang Lab at MIT identified 9 high quality sgRNAs (scores ≥ 50) with predicted Cas9-mediated cleavage within 35 bp of the mutation site. The guides are scored based on faithfulness of on-target activity, computed as 100% – weighted sum of offtarget hit-scores in the target genome. sgRNA 5 has the highest score at 82. Highlights, PAM sequences. Underlined, poly-T stretch. **(B)** Genomic locations of the unique PAM sequences corresponding to each sgRNA. The large red C represents the mutation site. **(C)** Number of nucleotides from predicted Cas9-mediated cleavage to the mutation site. sgRNA 4 is the closest, followed by sgRNA 3 and sgRNA 5. sgRNA 5, highlighted in red, is ultimately selected for targeted editing of *GATA1.*

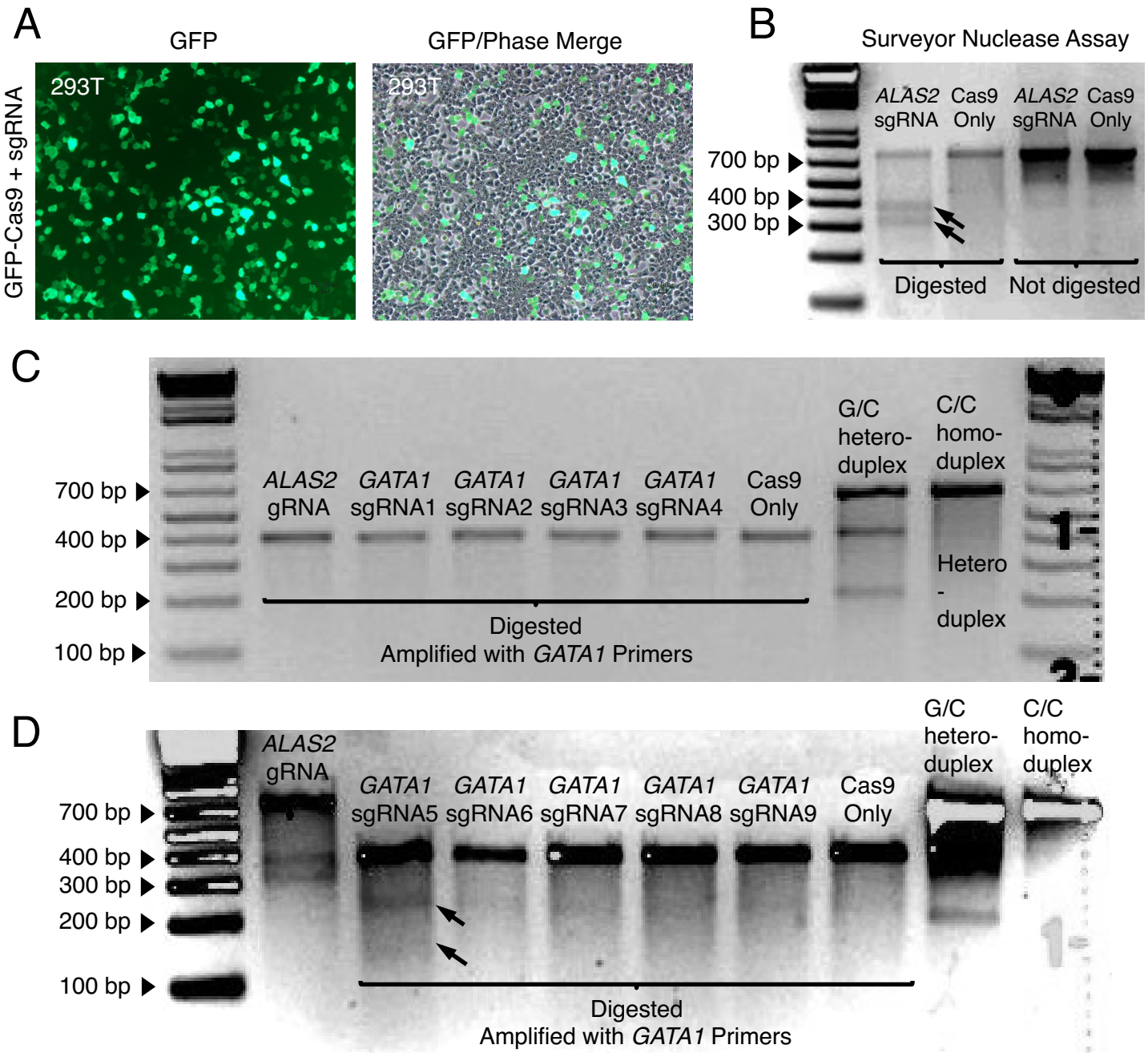


Figure 5. Surveyor nuclease assay identifies one guide RNA for targeted editing of *GATA1*
(A) 293T cells transfected with GFP-Cas9 and sgRNA. **(B)** Establishing the efficacy of the surveyor nuclease assay using an sgRNA known to produce targeted editing of *ALAS2*. This gene encodes delta-aminolevulinate synthase 2, which catalyzes the first step in the heme biosynthesis pathway. Genomic DNA harvested from 293T cells transfected with *ALAS2* sgRNA and Cas9 vs. Cas9 only are amplified with primers flanking the target region, heated to 95°C and cooled to 25°C to allow for DNA heteroduplex formation, and then treated with or without surveyor nuclease S, which cleaves DNA at mismatches. Ladder, 1 kb plus. PCR amplicon: 713 bp. Expected cleavage products based Cas9 cut site and indel (mismatch) formation: 387/326 bp. **(C-D)** Surveyor nuclease assay of *GATA1* sgRNA 1-9. PCR amplicon: 397 bp. Expected cleavage products: 259/138 bp (sgRNA1), 251/146 bp (sgRNA2), 247/150 bp (sgRNA3), 223/174 bp (sgRNA4), 250/147 bp (sgRNA5), 217/180 bp (sgRNA6), 199/198 bp (sgRNA7), 206/191 bp (sgRNA 8), and 217/180 bp (sgRNA9). Control G/C heteroduplex, assay positive control. PCR amplicon: 632 bp. Expected digested products: 416 and 217 bp. Control C homoduplex, assay negative control.

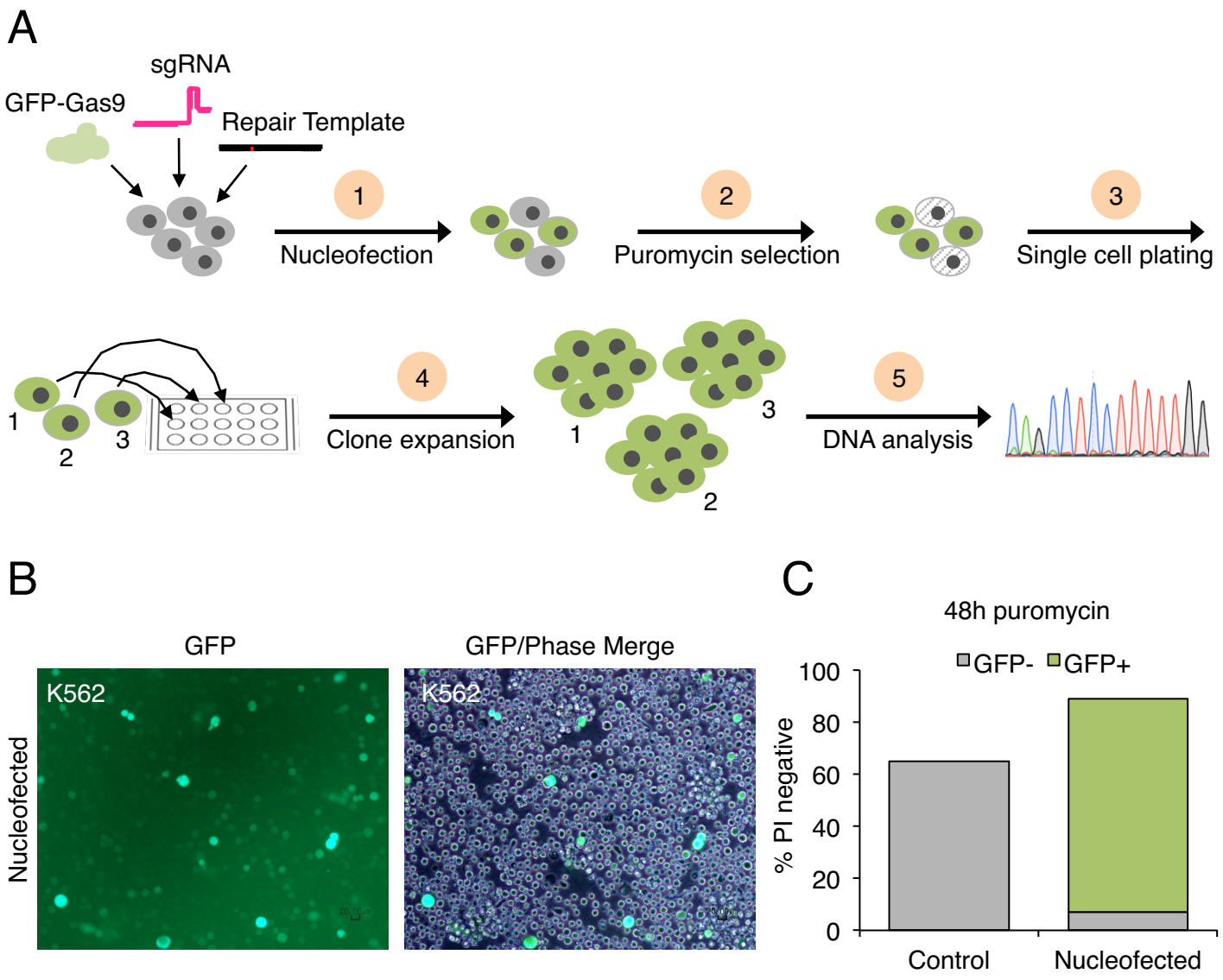


Figure 6. CRISPR-Cas9 offers the possibility of creating an isogenic human erythroid cell line with the mutation of interest in *GATA1*

(A) Schematic of targeted genome editing and generation of isogenic K562 cell lines. Step 1: nucleofect K562 cells with GFP-Cas9, sgRNA, and repair template with the single nucleotide mutation. Step 2: treat with puromycin to select for cells expressing GFP-Cas9. Step 3: dilute serially to distribute single cells in each well of a 96-well plate. Step 4: allow single cells to expand into observable clones under the microscope. Step 5: purify DNA from clones and screen for mutation of interest. **(B)** K562 cells nucleofected with GFP-Cas9, sgRNA, and repair template. **(C)** After 48 hours of puromycin, FACS analysis of 10,000 events showed that 65.1% of untreated (control) K562 cells and 88.9% of nucleofected K562 cells are propidium iodide (PI)-negative and 92.2% of nucleofected PI-negative K562 cells are GFP+. All untreated K562 cells are GFP-. The proportion of nucleofected PI-negative cells that are GFP+ is used to determine the appropriate serial dilution.

2.3 Increasing the ratio of DNA repair template to guide RNA to Cas9 dramatically improves *GATA1* editing efficiency and diversity

After identifying sgRNA 5 as a viable option for *GATA1* editing, I designed an antisense ssODN as the repair template to avoid Cas9-mediated cleavage directed by the guide's antisense targeting sequence. The ssODN contains adenine, corresponding to the thymine mutation, with flanking sequences of 90 bp homologous to the target region. I then nucleofected K562 cells with GFP-Cas9, antisense ssODN, and sgRNAs 5 and 6 (I included 6 to confirm correct interpretation of the surveyor nuclease assay and rule out possibility that faint digested bands were present indicating functional status). Of the 20 clones I generated from sgRNA 5, 4 (20%) showed indels at the Cas9 cleavage site: 1) a homozygous 1 bp insertion, 2) a homozygous 1 bp deletion, 3) a heterozygous 1 bp insertion and 1 bp deletion, and 4) a heterozygous 3 bp deletion (**Figure 7B**). In contrast, all 31 clones generated from sgRNA 6 showed wild-type sequence of the *GATA1* target region, suggesting that it is indeed a nonfunctional guide (**Figure 7B**).

Achieving targeted editing of *GATA1* with sgRNA 5 is an important milestone. In order to create a single nucleotide mutation 18 bp from the Cas9 cut site, however, I needed to increase the editing efficiency from 20% and create conditions that favor the less efficient HDR pathway. I hypothesized that increasing the mass ratio of DNA repair template to sgRNA to Cas9 in the nucleofection reaction could be a simple and effective approach. In my next experiment, I nucleofected 10:3:1 μg of ssODN to sgRNA 5 to Cas9 per million K562 cells, compared to the previous 2:2:1 ratio (**Figure 7C**). Remarkably, at this new ratio, 106 of 133 (79.7%) clones I isolated with identifiable sequence patterns demonstrated targeted editing of *GATA1* in at least one allele (**Figure**

7D). Furthermore, the diversity of *GATA1* modifications increased substantially. The range of indel sizes increased from 1 to 3 bp at the low ratio to 1 to 136 bp at the high ratio (**Figure 9-10**). Incorporation of the desired C-to-T mutation via HDR occurred in both alleles in 3 clones and in one allele in 2 clones (**Figure 11**). In the following sections, I will describe insights that emerged from characterization of all newly produced *GATA1* editing events.

2.4 Most *GATA1* edits occur within 15 bp of the Cas9 cleavage site and their distribution by genomic position reveals an upstream bias

Increased editing efficiency and diversity made many allelic combinations of *GATA1* modifications possible. It became much more difficult to identify the exact sequences of both alleles in each clone by visualization of Sanger sequencing data alone. To determine the size and genomic location of each indel, I used 3 computational tools that can read chromatograms and characterize insertions and deletions by comparing an experimental file to a control file. Tracking of Indels by DEcomposition (TIDE) reports percent contributions and associated p-values of insertions and deletions by size (Brinkman et al., 2014). CRISP-ID aligns possible indels of different sizes to a target genomic region (Dehairs et al., 2016). Inference of CRISPR Edits (ICE) combines features of both methods and provides percent contributions of indels by size *and* genomic location relative to the Cas9 cut site (Hsiau et al., 2018).

I analyzed *GATA1* editing in each of my experimental clones by comparing them to a control clone nucleofected with Cas9 and repair template but no sgRNA. I expanded a total of 157 clones for DNA analysis and was able to determine the indel sizes and

locations present in 133 clones. The remaining 24 clones had noisy data with inadequate alignment to the target sequence or unidentifiable indel. ICE was sufficient for analysis of 110 clones. For each sample, it generates a discordance plot that shows where the experimental and control sequence traces converge and diverge (**Figure 8A**). For a homozygous clone, ICE shows 1 predominant indel of a specific size and Sanger sequencing shows 1 trace (**Figure 8B**). For a heterozygous clone, ICE shows 2 predominant indels, one of which is often wild-type with 0 insertion or deletion, and Sanger sequencing shows 1 trace diverging into 2 traces in the edited region (**Figure 8C**). ICE also provides the genomic locations of the indels relative to the Cas9 cleavage site associated with sgRNA 5. In a homozygous clone, there is one major contributing sequence (**Figure 8D**). In a heterozygous clone, there are two major contributing sequences (**Figure 8E**). For the 23 clones that ICE failed to analyze, I was able to determine the indel sizes and locations using a combination of TIDE and CRISP-ID. Results from the 3 methods appear to correlate well for several high quality samples.

I mapped all modifications, including insertions, deletions, and substitutions, to the target region of *GATA1* by position relative to the Cas9 cleavage site (**Figure 9A**). Insertions ranged from 1 to 5 bp and deletions ranged from 1 to 136 bp. I presented the two alleles of each clone consecutively for ease of association. I then summed all insertions and deletions at each nucleotide and divided the numbers by total aligned sequences (n = 266 alleles from 133 clones) to determine the frequency of indel by genomic position (**Figure 9B**). Collectively, the data reveal that the majority of *GATA1* edits occurred within 15 bp of the Cas9 cleavage site. Editing at each nucleotide beyond –15 and +15 occurred at less than 3% frequency, except at the mutation site –18 from the

Cas9 cleavage site, which occurred at 5% frequency when the desired C-to-T substitutions are included with deletions.

Notably, the distribution of indels by genomic position reveals an upstream bias of the *GATA1* modifications. The editing frequency is higher upstream than downstream at each position equidistant from the Cas9 cleavage site (**Figure 9B**). Moreover, 3 unusually large homozygous deletions (26 bp from –29 to –4, 120 bp from –121 to –2, and 136 bp from –133 to +2) and 1 usually large heterozygous deletion (91 bp from –80 to +11) encompassed significant portions of intron 5 including the mutation site upstream of Cas9 cleavage. The boundary between intron 5 and exon 6 is located at +6-7 bp. One possible explanation for this upstream bias is that clones with homozygous indels downstream of the cut site have a survival disadvantage, as most of the downstream sequence is important for maintaining proper function of GATA1, via splicing regulation (e.g. the canonical splice acceptor site at +5-6) or translation into protein (e.g. the last exon begins at +7). Cells appear to tolerate upstream deletions of the intron better. Perhaps, proper splicing of *GATA1* can still occur with significant alterations of the sequence context as long as key nucleotides or motifs are conserved.

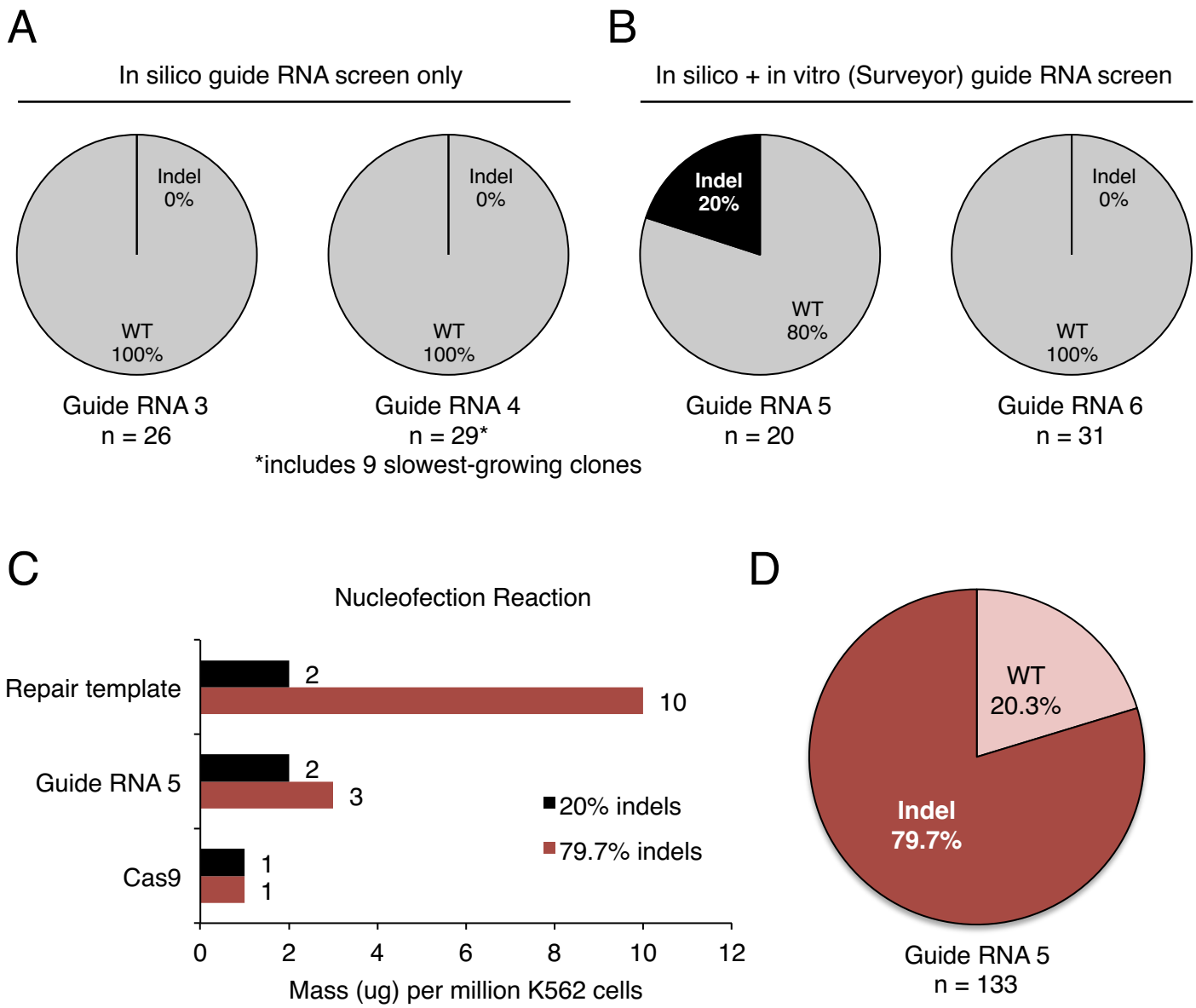


Figure 7. Rigorous screening of guide RNA and optimization of the nucleofection reaction increased *GATA1* editing efficiency from 0% to 79.7%

(A) No evidence of *GATA1* editing in over 50 K562 clones generated with guide RNA 3 and 4 identified *in silico*, GFP-Cas9, and a repair template with the mutation of interest. Selective screening of slow-growing clones did not reveal any indels or single nucleotide replacement in the expected region. **(B)** Guide RNA 5 identified *in silico* and *in vitro* with the surveyor nuclease assay produced K562 clones with targeted *GATA1* editing. Of the 20 clones screened, 4 had indels at the Cas9 cleavage site: 1) a homozygous 1 bp insertion, 2) a homozygous 1 bp deletion, 3) a heterozygous 1 bp insertion and 1 bp deletion, and 4) a heterozygous 3 bp deletion. In contrast to sgRNA 5, sgRNA 6 did not show definite activity in the surveyor nuclease assay. CRISPR using this guide did not produce any K562 clones with evidence of *GATA1* editing. **(C-D)** Increasing the mass ratio of repair template to sgRNA to Cas9 from 2:2:1 to 10:3:1 in the nucleofection reaction dramatically improved *GATA1* editing efficiency from 20% to 79.7%. The range of indel sizes also increased from 3 bp to over 100 bp. See Figures 9 and 10 for characterization of all editing events produced in these optimized conditions.

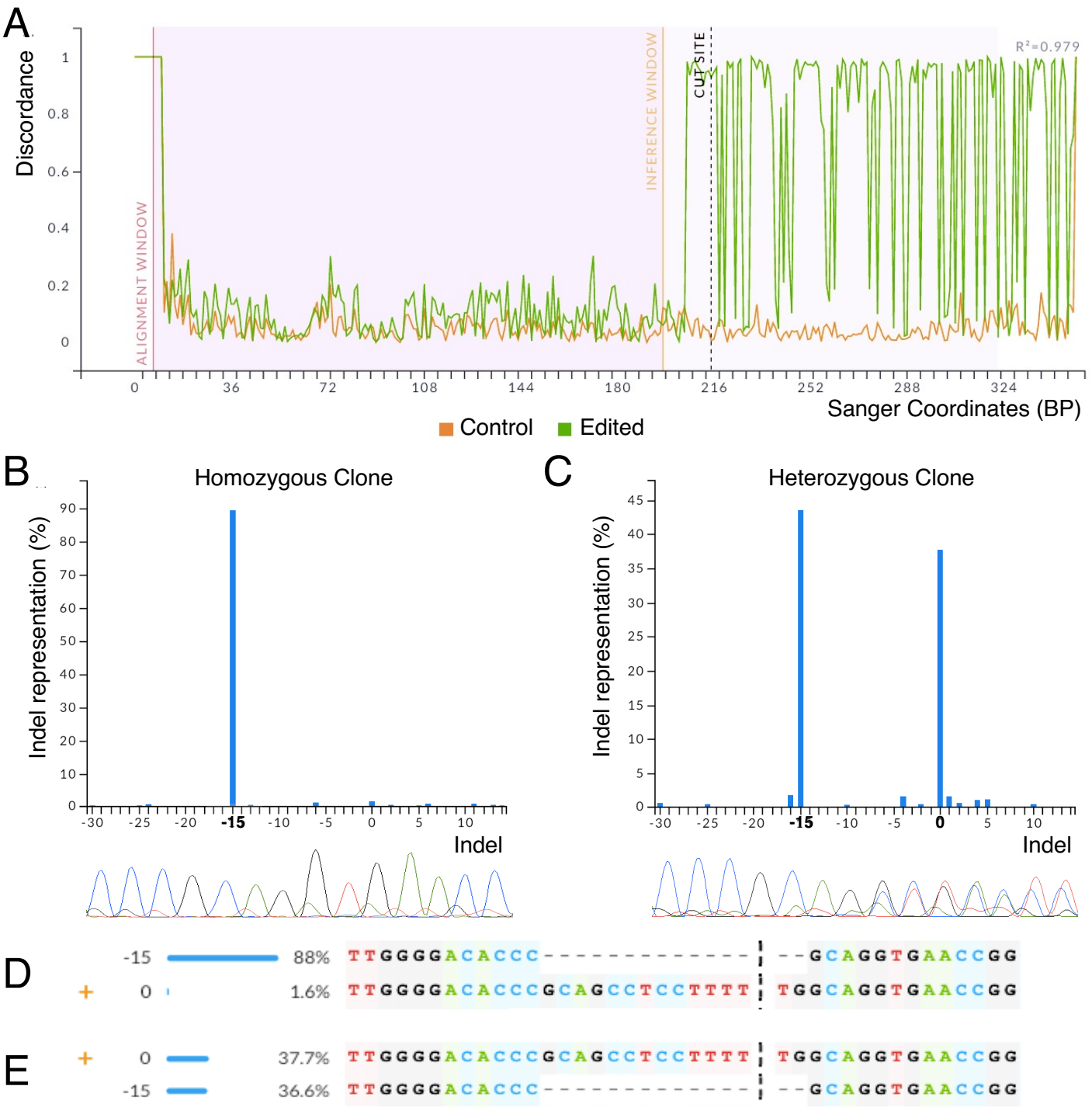


Figure 8. Inference of CRISPR Edits (ICE) identifies allele-specific indels in isogenic cell lines
(A) Discordance plot generated by ICE (Hsiau et al., 2018). Control (orange tracing), Sanger sequencing of a K562 clone nucleofected with GFP-Cas9 and repair template but no sgRNA. Edited (green tracing), Sanger sequencing of a K562 clone nucleofected by sgRNA 5, GFP-Cas9, and repair template. The two traces are closely aligned until 10-20 bp from the predicted Cas9 cut site. **(B)** Homozygous clone. 15 bp deletion (-15) is the only indel present. Sanger sequencing shows 1 trace. **(C)** Heterozygous clone. Both 15 bp deletion (-15) and wild-type sequence (0) are present. Sanger sequencing shows 1 trace diverging into 2 traces in the edited region. **(D)** The homozygous clone has 1 significant contributing sequence. **(E)** The heterozygous clone has 2 significant contributing sequences. Horizontal dash, deleted nucleotides. Vertical dash, Cas9 cut site. Plus sign, wild-type sequence.

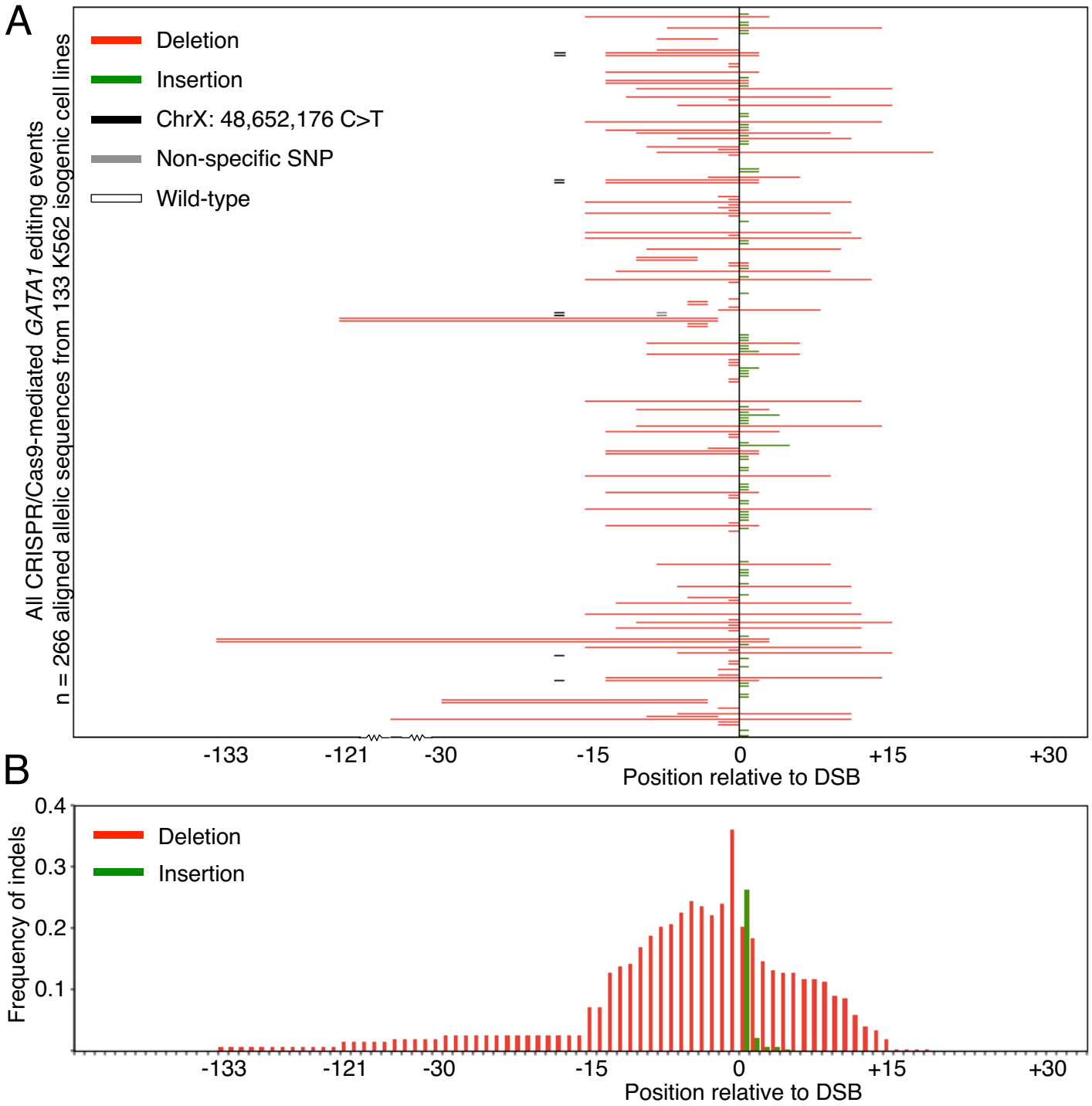


Figure 9. Most *GATA1* edits occur within 15 bp of the Cas9 cleavage site and their distribution by genomic position reveals an upstream bias

(A) Distribution of indel location. Each line represents the aligned sequence of an allele identified by ICE or a combination of TIDE and CRISP-ID (Brinkman et al., 2014, Dehairs et al., 2016). The two alleles of each K562 clone are mapped consecutively. Of note, encompassing the mutation site are 3 unusually large homozygous deletions (26 bp from -29 to -4; 120 bp from -121 to -2; 136 bp from -133 to +2) and 1 usually large heterozygous deletion (91 bp from -80 to +11). **(B)** Frequency of indel by position. Editing of the mutation site at -18 from DSB, including deletion (red) and the desired C>T substitution (not represented), occurred at 5% frequency. The intron-exon boundary is between positions +6 and +7. The editing frequency is higher upstream than downstream at each position equidistant from the DSB.

2.5 Most *GATA1* edits occur within the intron and those that alter the canonical splice acceptor site or adjacent exon are exclusively heterozygous

Consistent with upstream bias of *GATA1* modifications, the majority (70%) of edited clones contain indels that map exclusively to intron 5 (**Figure 10B**). Even though exon 6 begins 7 bp from the Cas9 cleavage site, only a minority (30%) of edited clones contains indels that span both intron 5 and exon 6. Importantly, over a third (36%) of edited clones demonstrate homozygous editing of *GATA1* (**Figure 10C**). However, all clones with deletion of the canonical splice acceptor site or exon 6 are heterozygous (n = 35) (**Figure 10D**). These results further suggest that proper splicing and translation of GATA1, at least from one allele, is required for K562 cell survival.

The distribution of *GATA1* indels by size shows that the most frequent modifications are 1 bp insertions followed by 1-2 bp deletions at the Cas9 cleavage site 6 bp upstream of the intron 5-exon 6 boundary (**Figure 10A**). Curiously, the next most frequent editing event is a 15 bp intronic deletion –13 to +2 bp relative to the Cas9 cleavage site. Its consistent genomic location and frequency comparable to 2 bp deletions argue against a completely stochastic process. Given its presence in both homozygous and heterozygous clones, this defined region does not appear to be essential for cell survival and may in fact be selected for deletion. Further inspection reveals that this intronic deletion encompasses a potential alternative splice acceptor site (YAG, where Y is a pyrimidine, or C in this case) as well as the PAM sequence of sgRNA 5 (CCT in the sense strand corresponds to AGG in the antisense strand detected by Cas9) and the first 5 bp of its 3' seed sequence.

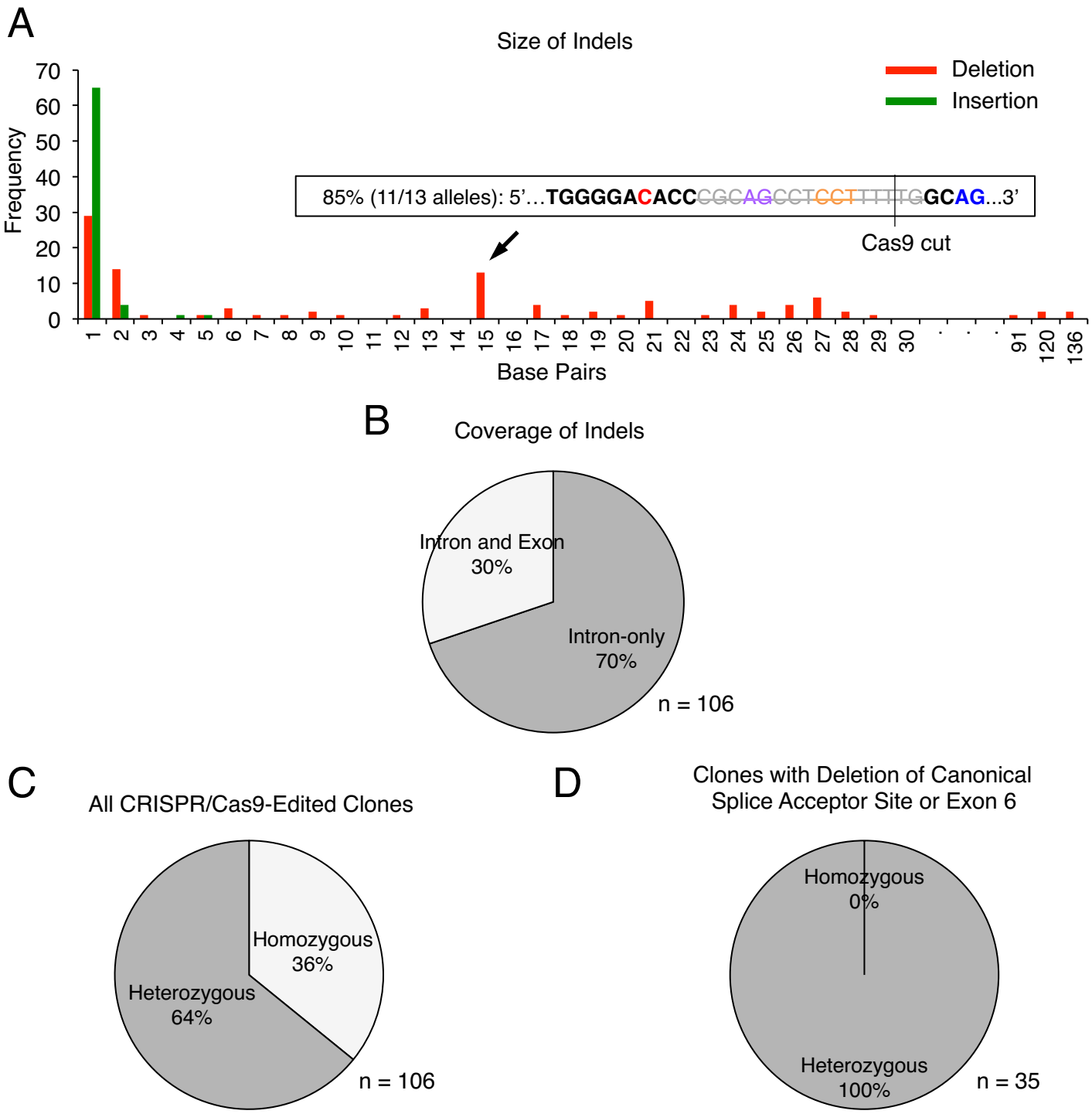


Figure 10. Most *GATA1* edits occur within the intron and those that alter the canonical splice acceptor site or adjacent exon are exclusively heterozygous

(A) Distribution of indel size. Insertions range from 1 to 5 bp. Deletions range from 1 to 136 bp. Other than indels of 1-2 bp, a 15 bp deletion is the most frequently observed editing event. Most of the 15 bp deletions span -13 to +2 relative to the Cas9 cleavage site, which encompass an alternative splice acceptor site (purple) and sgRNA 5 PAM sequence (orange) but leave the canonical splice acceptor site (blue) intact. **(B)** The majority of indels are localized within the intron, even though exon 6 is only 7 bp from the Cas9 cleavage site. **(C)** Over a third of all edited clones have homozygous indels. **(D)** All clones with deletion of the canonical splice acceptor site or the adjacent exon are heterozygous.

2.6 The desired C-to-T intronic mutation in *GATA1* co-occurs with additional modifications such as deletion of the guide RNA PAM sequence and a potential alternative splice acceptor site

Among the 133 clones with identifiable sequence patterns, I found 3 clones with homozygous C-to-T mutation (Mutants 1-3) and 2 clones with heterozygous C-to-T mutation (Mutants 4-5) at the desired genomic position (**Figure 11A**). In addition to substitution, editing at this position includes 3 unusually large homozygous deletions ranging from 26 to 136 bp and 1 unusually large heterozygous 91 bp deletion described previously (**Figure 9A**, **Figure 11B**). Interestingly, the homozygous C-to-T mutation co-occurred in Mutants 1-2 with the frequently observed 15 bp deletion –13 to +2 bp from the Cas9 cleavage site, spanning a potential alternative splice acceptor site and the sgRNA 5 PAM and 3' seed sequences (**Figure 11C**). It co-occurred in Mutant 3 with a novel C-to-A homozygous mutation between the potential alternative splice acceptor site and the PAM sequence. The heterozygous C-to-T mutation could be located on either allele so it is difficult to identify its associated downstream modifications. No additional modification is possible when an otherwise wild-type allele is present, as in Mutant 4.

Both technical and biological reasons can explain the presence of additional modifications in isogenic cell lines that have recapitulated the desired C-to-T mutation in *GATA1*. From a technical perspective, the error-prone NHEJ pathway occurs much more frequently than the high-fidelity HDR pathway, which is generally active only in dividing cells (Saleh-Gohari and Helleday 2004). Alleles that incorporated the mutation via HDR can undergo repeated editing by CRISPR-Cas9 until the target locus is rendered unrecognizable to sgRNA via indels from NHEJ. Concurrent deletion of the sgRNA 5

PAM sequence and 3' seed sequence essential for target binding and Cas9 cleavage supports this phenomenon. One may be able to overcome this obstacle by enhancing the efficiency of HDR or blocking recurrent RNA-guided Cas9 editing. Recent approaches to enhancing HDR include cell cycle synchronization (Lin et al., 2014, Yang et al., 2016) and ssODN optimization with asymmetric 5' and 3' homology arms (Richardson et al., 2016). To isolate a desired mutation and block additional modifications, Kwart et al. (2017) developed a framework named CORRECT consisting of two rounds of genome editing (Kwart et al., 2017). During the first round, the desired mutation is generated along with a CRISPR/Cas-blocking mutation positioned in the PAM sequence (re-Cas) or 3' seed sequence of the target sgRNA (re-Guide) using a repair template with both mutations. During the second round, the CRISPR/Cas-blocking mutation is erased using VRER-Cas9 (re-Cas), a Cas9 variant that recognizes the previously altered PAM sequence, or a modified sgRNA (re-Guide) that recognizes the previously altered 3' seed sequence, and a repair template with the desired mutation only (Kleinstiver et al. 2015). The authors performed CORRECT in human pluripotent stem cells and estimate that generation of scarlessly edited isogenic cell lines requires approximately 3 months.

From a biological perspective, it possible that isogenic cell lines with the homozygous C-to-T mutation are at a survival disadvantage and additional modifications are required to "rescue" this defect. I hypothesized that this intronic mutation causes the distinct dyserythropoietic anemia observed in our patients via altered splicing of GATA1 leading to reduced level of the functional protein. Remarkably, both Mutant 1 and Mutant 2 with the desired homozygous mutation have an accompanying 15 bp deletion –13 to +2 bp from the Cas9 cleavage site. The consistency of this genomic location suggests that

they are not random events as one might expect with indels from NHEJ. These 15 deletions include 6 bp upstream of the sgRNA 5 PAM sequence and 3' seed sequence not involved in targeting of CRISPR-Cas9. Interestingly, present in this 6 bp region is a potential alternative splice acceptor site (CAG). The C-to-T mutation may cause altered splicing and loss of function of GATA1 via aberrant activation of this alternative splice acceptor site and a partial intron retention event. Previously, I have discussed that K562 cells may require at least one functional copy of GATA1 for survival, given an upstream intronic bias of all targeted editing events and lack of homozygous clones with alteration of the canonical splice acceptor site or adjacent exon (**Figures 9-10**). If the desired homozygous C-to-T mutation renders both copies of GATA1 nonfunctional and K562 cells indeed require some level of functional GATA1 for survival, then it might be biologically impossible to isolate isogenic cell lines with only the desired mutation. The additional modifications may rescue the splicing defect via deletion of the potential alternative splice acceptor site (Mutants 1-2) or spontaneous mutation of a nearby nucleotide (Mutant 3) to preserve proper splicing and function of GATA1.

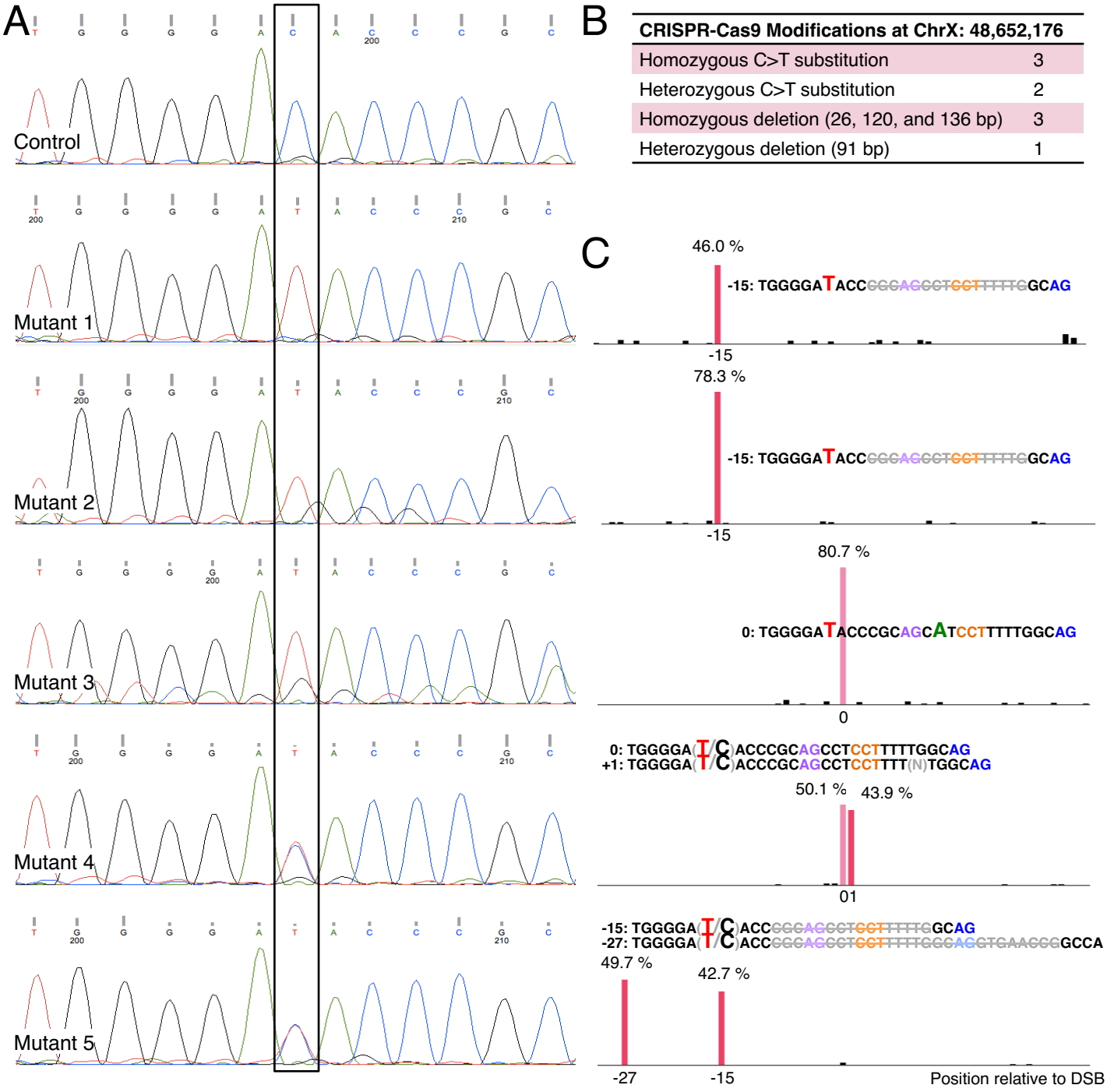


Figure 11. The desired C-to-T intronic mutation in *GATA1* co-occurs with additional modifications including deletion of the guide RNA PAM sequence and a potential alternative splice acceptor site
(A) Sanger sequencing of K562 clones with the desired C>T mutation in the last intron of *GATA1*. Control, nucleofected with GFP-Cas9 and and repair template but no sgRNA. Mutants 1-3, homozygous T/T. Mutants 4-5, heterozygous C/T. **(B)** All types of editing at the desired genomic location. The 4 clones with large deletions spanning this position are described in Figure 9A. **(C)** TIDE analysis of Mutants 1-5. Homozygous C>T mutation co-occurred in 2 clones with a 15 bp deletion including the sgRNA 5 PAM sequence and a potential alternative splice acceptor site and in 1 clone with a C>A mutation 2 bp from the potential alternative splice acceptor site. Pink bars, significant indels ($p < 0.001$) and its % representation. Black bars, insignificant indels ($p \geq 0.001$). Modified sequences: strikethrough, deleted; blue, canonical splice acceptor site; purple, alternative splice acceptor site; orange, sgRNA 5 PAM sequence; green, SNP.

METHODS

Cloning of single guide RNA constructs

The human *GATA1* gene sequence was obtained from NCBI and a 100-bp region centering on the mutation site of interest (chrX: 48,652,176) was submitted to Optimized CRISPR Design developed by the Feng Zhang laboratory at MIT for identification of appropriate sgRNAs. Targeting sequences with computed scores ≥50 were considered high quality based on faithfulness of on-target activity. Nine sgRNAs with the highest scores and the shortest distance from the predicted Cas9 cleavage site to the mutation site were selected as candidates for experimentation (**Figure 4**). Two sets of oligonucleotides were purchased for each sgRNA: 1) the top strand of contains GTTTT 3' overhang complementary to CAAAA in the linearized vector and constitute the first 5 bases of tracrRNA; 2) the bottom strand contains CGGTC 3' overhang complementary to CACCG in the linearized vector and constitutes the last 4 bases of the U6 promoter and the first base required for PolIII transcription start site. For cloning into the pSg1 plasmid vector (GenScript), 0.5 μg of the plasmid was digested with 5U BplI in 1X Tango Buffer supplemented with 0.05 mM S-adenosylmethionine (SAM) for 2 hours at 37°C, purified with QiaQuick columns (Qiagen), re-digested with 5U NheI in NEB Buffer 2.1 for 1 hour at 37°C, dephosphorylated with FastAP thermosensitive alkaline phosphatase (Thermo Fisher) for 15 min. at 37°C, and then purified again with QiaQuick columns (Qiagen). Meanwhile, the sgRNA oligos were each diluted to 20 μM in 5 μl water, heated to 85°C, snap cooled on ice, and then phosphorylated with T4 polynucleotide kinase (NEB) at 37°C for 30-60 min. The phosphorylated top and bottom oligos were mixed and then denatured and annealed using the following temperature series: 98°C for 2 min., 85°C for

2 min., 75°C for 5 min., 65°C for 5 min., and room temperature. The digested pSg1 plasmid backbone and phosphorylated and annealed oligos were ligated with T4 DNA ligase (NEB) at room temperature for 15 min. For transformation, each ligation reaction was added to 10 µl of chemically competent *E. coli* (Invitrogen™), heat shocked at 42°C for 40 seconds, and then returned to ice. Transformed bacteria were shaken in 500 µl LB media at 37°C for 45-60 min. and then plated on LB agar plate with ampicillin overnight. The next day, colonies were inoculated in 3 mL ampicillin-supplemented LB media and shaken at 37°C for 24 hours. Plasmid DNA was isolated by miniprep (Qiagen) and sequenced for identification of integrated sgRNA*. For mass production of plasmids carrying sgRNA, clones confirmed by sequencing were expanded for maxiprep (Qiagen).

*Sequencing primer:

U6 forward primer: 5'-TGTACAAAAAAGCAGGCTTTAAAGG -3'

Functional validation of single guide RNA constructs

293T cells were cultured in DMEM (Gibco™) with 10% FBS and 1% penicillin-streptomycin. For functional validation of each sgRNA construct, 300,000 cells were transfected with 0.66 µg sgRNA (cloned pSg1 plasmid) and 0.33 µg GFP-Cas9 (pXPR_001 plasmid from Feng Zhang) in Opti-MEM (Thermo Fisher) with FuGENE (Promega). 24 and 48 hours after transfection, puromycin was added to fresh media at a final concentration of 2 µg/ml for selection of cells that have been successfully transfected. 72 hours after transfection, the cells were harvested for genomic DNA extraction (Qiagen). For the surveyor nuclease assay (IDT), the target genomic region in *GATA1* was first amplified by PCR*, along Control G and Control C plasmids with

premixed primers. After purification, DNA heteroduplexes were generated using the following temperature series: 95°C for 10 min., -2.0°C per second from 95°C to 85°C, and then -0.3°C per second from 85°C to 85°C to 25°C. The entire cooling process took about 20 min. total. Equal amounts of Control C and Control G were mixed for the assay positive control. Only Control C was present in the assay negative control. The annealed heteroduplexes were then digested with Surveyor Nuclease S and Surveyor Enhancer S in water with 0.15M $MgCl_2$ at 42°C for 60 min. The digested products were visualized on 2% agarose gel. Electrophoresis was stopped after the dye has migrated about 1/3 of the length of the gel for the purpose of photographing the gel early in the run. Then electrophoresis was continued until the dye has migrated 2/3 of the length of the gel.

*PCR primers:

GATA1 forward primer: 5'-CCTCCTCCTTCCTCTCCTCT-3'

GATA1 reverse primer: 5' CACCACCATAAAGCCACCAG-3'

Genome editing and generation of K562 isogenic cell lines

K562 cells were cultured in RPMI 1640 medium (Gibco™) with 10% FBS and 1% pencillin-streptomycin. On the day prior to nucleofection, they were plated at a density of 250,000 cells/mL without antibiotics. For genome editing with CRISPR/Cas9, 1 million cells were resuspended in Supplemented Nucleofector® Solution (Lonza) with either 2 μg DNA repair template (single-stranded oligo donor with intended mutation flanked by 90-bp homology arms), 2 μg sgRNA, and 1 μg GFP-Cas9 (*low ratio*) OR 10 μg DNA repair template, 3 μg sgRNA, and 1 μg GFP-Cas9 (*high ratio*). The cell suspensions were then transferred to a certified cuvette and placed in a standard device

for nucleofection (Nucleofector® Program T-016 for K562 cells). After nucleofection, the cells were transferred back to its normal growth media and incubated in a humidified 37°C/5% CO_2 chamber. 24 and 48 hours after nucleofection, puromycin was added to fresh media at a final concentration of 2 μg/ml for selection of cells that have been successfully nucleofected. 72 hours after nucleofection, the cells were resuspended in media without puromycin. 96 hours after nucleofection, 300 μL of each sample was collected, centrifuged, and resuspended in FACS buffer with propidium iodide (PI). FACS was performed to determine the percentage of PI-negative cells (indicating survival) that were GFP+ (indicating Cas9 expression), which informed subsequent serial dilutions, i.e. if there were 100,000 cells initially and 60% PI-negative cells are GFP+, I used 60,000 as the starting number. Cell suspensions were serially diluted from the starting number to 50,000/mL, 5,000/mL, 500/mL, and 10/mL to achieve single cell plating. The 96-well plates were then incubated for 10-14 days and screened under light microscope for identification of well-defined single colonies in each well. All equivocal single colonies and double or triple colonies were excluded from further analysis. Single colonies were then transferred to larger plates for clonal expansion.

Screening for GATA1 mutation and DNA sequencing analysis

Expanded clones were harvested for genomic DNA extraction (Qiagen) followed by PCR amplification and sequencing of the *GATA1* target region*. Sanger sequencing chromatograms were analyzed using ICE (Hsiau et al., 2018). If ICE failed to align the sequences, a combination of TIDE and CRISP-ID were used to identify allele-specific alterations (Brinkman et al., 2014; Dehairs et al., 2016). The control clone used in these

methods was nucleofected with GFP-Cas9 and DNA repair template but no sgRNA. All modifications were mapped by genomic position on Microsoft Excel and editing frequencies at each position was determine by dividing the total number of editing events at that position by total number of alleles.

*PCR primers:

GATA1 forward primer: 5'-CCTCCTCCTTCCTCTCCTCT-3' (also sequencing primer)

GATA1 reverse primer: 5' CACCACCATAAAGCCACCAG-3'

REFERENCES

Brinkman, E.K., Chen, T., Amendola, M., and van Steensel, B. (2014). Easy quantitative assessment of genome editing by sequence trace decomposition. Nucleic Acids Res. *42*, e168.

Chen, F., Pruett-Miller, S.M., Huang, Y., Gjoka, M., Duda, K., Taunton, J., Collingwood, T.N., Frodin, M., and Davis, G.D. (2011). High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. Nat. Methods *8*, 753–755.

Dehairs, J., Talebi, A., Cherifi, Y., and Swinnen, J.V. (2016). CRISP-ID: decoding CRISPR mediated indels by Sanger sequencing. Sci Rep *6*, 28973.

Hsiau, T., Maures, T., Waite, K., Yang, J., Kelso, R., Holden, K., and Stoner, R. (2018). Inference of CRISPR Edits from Sanger Trace Data. bioRxiv.

Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Topkar, V.V., Nguyen, N.T., Zheng, Z., Gonzales, A.P.W., Li, Z., Peterson, R.T., Yeh, J.-R.J., et al. (2015). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature *523*, 481–485.

Kwart, D., Paquet, D., Teo, S., and Tessier-Lavigne, M. (2017). Precise and efficient scarless genome editing in stem cells using CORRECT. Nat Protoc *12*, 329–354.

Lin, S., Staahl, B.T., Alla, R.K., and Doudna, J.A. (2014). Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. Elife *3*, e04766.

Lozzio, C.B., and Lozzio, B.B. (1975). Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. Blood *45*, 321–334.

Naumann, S., Reutzel, D., Speicher, M., and Decker, H.J. (2001). Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization. Leuk. Res. *25*, 313–322.

Nielsen, S., Yuzenkova, Y., and Zenkin, N. (2013). Mechanism of eukaryotic RNA polymerase III transcription termination. Science *340*, 1577–1580.

Qiu, P., Shandilya, H., D'Alessio, J.M., O'Connor, K., Durocher, J., and Gerard, G.F. (2004). Mutation detection using Surveyor nuclease. BioTechniques *36*, 702–707.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat Protoc *8*, 2281–2308.

Richardson, C.D., Ray, G.J., DeWitt, M.A., Curie, G.L., and Corn, J.E. (2016). Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. Nat. Biotechnol. *34*, 339–344.

Saleh-Gohari, N., and Helleday, T. (2004). Conservative homologous recombination preferentially repairs DNA double-strand breaks in the S phase of the cell cycle in human cells. Nucleic Acids Res. *32*, 3683–3688.

Song, F., and Stieger, K. (2017). Optimizing the DNA Donor Template for Homology-Directed Repair of Double-Strand Breaks. Mol Ther Nucleic Acids *7*, 53–60.

Ulirsch, J.C., Lacy, J.N., An, X., Mohandas, N., Mikkelsen, T.S., and Sankaran, V.G. (2014). Altered chromatin occupancy of master regulators underlies evolutionary divergence in the transcriptional landscape of erythroid differentiation. PLoS Genet. *10*, e1004890.

Yang, D., Scavuzzo, M.A., Chmielowiec, J., Sharp, R., Bajic, A., and Borowiak, M. (2016). Enrichment of G2/M cell cycle phase in human pluripotent stem cells enhances HDR-mediated gene repair with customizable endonucleases. Sci Rep *6*, 21264.

CHAPTER 3. Impaired human hematopoiesis due to a cryptic intronic *GATA1* splicing mutation*

**Adapted from paper in preparation for submission (some components are incomplete)*

Nour J. Abdulhay[1,2], Jeffrey M. Verboon[1,2], Jacob C. Ulirsch[1,2], Barbara Zieger[3], Xiaoli Mi[1,2], Esther A. Obeng[1,2,4], Miriam Erlacher[3], Namrata Gupta[2], Stacey B. Gabriel[2], Benjamin L. Ebert[2,4], Charlotte Niemeyer[3], Rami N. Khoriaty[5], Philip Ancliff[5], Hanna T. Gazda[2,6], Marcin W. Wlodarski[3], and Vijay G. Sankaran[1,2]

[1]Divison of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA.

[2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

[3]Division of Pediatric Hematology and Oncology, Department of Pediatrics and Adolescent Medicine, Faculty of Medicine, Medical Center-University of Freiburg, Freiburg, Germany.

[4]Division of Hematology, Brigham and Women's Hospital, Boston, Massachusetts, USA.

[5]Department of Paediatric Haematology, Great Ormond Street Hospital for Children, Great Ormond Street, London, United Kingdom.

[6]Division of Hematology and Oncology, Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA.

[7]Division of Genetics and Genomics, The Manton Center for Orphan Disease Research, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA.

ABSTRACT

Recent studies of allelic variation underlying genetic blood disorders have provided important insights into human hematopoiesis. In the majority of instances, such pathogenic mutations result in complete loss-of-function or missense changes. It can be challenging to assess the pathogenicity of non-coding variants. Here, we characterize two unrelated patients with a distinct presentation of dyserythropoietic anemia and disordered hematopoiesis associated with an intronic mutation in *GATA1* that is 24 nucleotides upstream of the canonical splice acceptor site. Functional studies demonstrate that this single nucleotide alteration leads to reduced canonical splicing and increased use of an alternative splice acceptor site that causes a partial intron retention event. The resultant altered GATA1 protein due to this intron retention event – involving a 5 amino acid insertion between the two zinc fingers - has no observable function. Collectively our results demonstrate how altered splicing of GATA1, which leads to overall reduced levels of the normal form of this master hematopoietic transcription factor, can result in distinct alterations in hematopoiesis, with potential implications for more common acquired blood disorders due to splicing factor mutations, such as the myelodysplastic syndromes.

INTRODUCTION

While hematopoiesis is arguably one of the best-understood paradigms of cell differentiation in physiology, many facets of this process remain poorly understood. Genetic blood disorders provide an opportunity to learn more about hematopoiesis, even in cases where a mutated gene has been previously well studied, since allelic variation can provide novel biological insights (1-5). Advances in sequencing technologies have enabled rapid and efficient mutation identification, but deciphering the pathogenicity of non-coding variation can be immensely challenging. Non-coding transcriptional regulatory elements can be interrogated using a number of functional approaches, including the use of genome editing and exogenous assays of regulatory function (6-8). However, deciphering the pathogenicity of cryptic splicing mutations can present substantial challenges (9, 10).

Here we have identified two unrelated patients with a unique form of dyserythropoietic anemia that is associated with other abnormalities in hematopoiesis. Both of the patients harbor identical intronic mutations in *GATA1*. Through functional studies, we show that this single nucleotide change leads to altered splice acceptor usage and overall decreased splicing efficiency, thereby resulting in the disease phenotype. This mutation disrupts the activity of the U2 splicing complex in the last intron of *GATA1*, providing a potential connection between these rare forms of impaired hematopoiesis and more commonly observed cases of myelodysplastic syndrome (MDS), which are attributable to somatic mutations in this splicing regulatory complex itself (11). In addition, our study illustrates how decreasing the splicing efficiency of the hematopoietic master regulator GATA1 can impair selective aspects of human hematopoiesis and lead

to distinct phenotypes when compared with other pathogenic mutations affecting the same gene (12-14).

RESULTS & DISCUSSION

In the course of studying a cohort of patients with rare genetic blood disorders, we encountered two patients who both had a dyserythropoietic anemia that necessitated transfusions in early infancy with subsequent evolution into a milder anemia that was noted to worsen in the setting of intercurrent illnesses. Bone marrow evaluation of the patients revealed dyserythropoiesis, along with frequent small hypolobulated megakaryocytes and occasional dysplastic myeloid cells (**Figure 1A, Supplemental Figure 1**). Platelet function testing in Patient 1, who was noted to have clinical bleeding, revealed specific impairments, including defective aggregation and α-/δ-granule secretion (as indicated by decreased expression of P-selectin/ CD62 and LAMP-3/ CD63), which are suggestive of defective thrombopoiesis (**Figure 1B and Supplemental Figure 1**). Further clinical details of the 2 patients are described in the Methods and in the supplemental material. To assess the etiology of the impaired hematopoiesis in these patients, whole exome sequencing (WES) and targeted mutation analysis were performed. No known blood disorder-associated mutations were identified, but both patients were noted to have a unique mutation in the fifth intron of *GATA1* (ChrX:48,652,176 C>T in hg19) that was carried in the mothers of these patients (**Figure 1C**, **Supplemental Figure 2**). This mutation was absent from the 123,136 exomes and 15,496 genomes in the gnomAD database, despite excellent coverage of this region (**Supplemental Methods**).

The identified mutation was located 24 nucleotides upstream of the canonical splice acceptor site in the fifth intron of *GATA1*. While the mutation was not predicted to disrupt known splicing elements, we hypothesized that it may impair appropriate splicing of this gene. To directly interrogate this, we utilized a minigene reporter assay to examine how this single nucleotide alteration may impact splicing (**Figures 2A, B**) (15). We found that the mutation reduced normal splicing of this region of *GATA1* and promoted an intron retention event of 15 nucleotides involving an alternative splice acceptor site (**Figure 2B**). Semi-quantitative RT-PCR analysis revealed a reduction of canonical *GATA1* splicing to 42% of normal levels, with the mutant mRNA at 36% of normal levels (**Figure 2C**). Importantly, we could confirm this altered splicing was present in the patient samples, but was found at significantly lower levels in healthy controls (**Figure 2D**). Throughout human erythroid differentiation (7, 16), there was no major variation in GATA1 splicing that could be identified (**Figure 2E**, **Supplemental Figure 3**), suggesting that the altered splicing observed in these cases was attributable to the distinctive pathogenic variants.

We wanted to further investigate the underlying mechanisms for the altered splicing present due to these pathogenic mutations. It was interesting that the observed alternative splice acceptor site usage with resultant intron retention (**Figure 2F**) was similar to what is frequently observed in *SF3B1*-mutated MDS, where there is alternative splice acceptor site usage in a number of transcripts between approximately 15 and 25 nucleotides upstream of the canonical splice acceptor site (17, 18). Given the phenotypic similarity between MDS and the disordered hematopoiesis observed in these patients, as well as the likely involvement of the U2 splicing complex in the region where this

mutation resided, we decided to test whether the commonly observed SF3B1 K700E mutant, which is known to impair U2 splicing activity, would further impair the observed altered splicing due to this intronic *GATA1* mutation. Using the minigene assay described above, we found that expression of the SF3B1 K700E mutant markedly impaired splicing of the mutant *GATA1* and led to near complete alternative splice acceptor usage. This finding shows that the observed germline mutation in these patients directly impairs activity of the U2 splicing complex in this region of GATA1. We examined whether SF3B1-mutated MDS cases (with a diagnosis of refractory anemia with ringed sideroblasts) may have altered splicing or levels of *GATA1*, but we did not observe consistent defects suggesting altered splicing or the presence of reduced levels of *GATA1* mRNA, even in sorted stage-matched bone marrow populations (**Supplemental Figure 4**). However, it is possible that a potential defect may be present and detection of this may be masked due to selective survival of or selection for cells that have normal *GATA1* splicing.

To gain further insight into the mechanism by which the *GATA1* intronic mutant disrupts human hematopoiesis, we examined whether the protein produced from the intron retention event – which would add 5 additional amino acids between the N and C-terminal zinc fingers of GATA1 – may have altered activity (**Figure 3A**). In an exogenous setting, we found that both RNA levels (as inferred through the expression of a linked GFP molecule that is translated from an internal ribosome entry site on the same transcript) and protein levels (as measured directly) of this mutated cDNA were stable and similar to what is observed with wild type GATA1 cDNA (**Figure 3B**). Exogenous expression of the mutated cDNA in primary human hematopoietic stem and progenitor

cells (HSPCs) induced to undergo erythroid differentiation revealed that the mutated form failed to promote precocious erythroid differentiation, as occurs with the wild type cDNA, but concomitantly there was no dominant negative activity observed in this setting where wild type GATA1 protein is present (**Figure 3C**) (6, 19, 20). These results suggest that the mutated protein formed by this intron retention event is inactive. To directly assess this, we used a *Gata1*-null mouse cell line, G1E, to complement this phenotype with either the wild type or mutant cDNAs (14, 21). While the wild type cDNA robustly promoted erythroid differentiation in this setting, the mutant form failed to do so, directly showing that this mutated protein produced from an intron retention event displays complete loss-of-function (**Figure 3D**, **Supplemental Figure 5**). Together with the data shown above from the interrogation of splicing, these results demonstrate that the impaired hematopoiesis in these cases emerges due to reduced expression of GATA1 to ~40% of normal levels with the production of an inactive mutant protein due to the intron retention event we observed.

Our findings have several important and broad implications. While cryptic splicing mutations can be challenging to study, we illustrate how by using a series of functional assays, the mechanisms underlying such mutations can be more fully understood. Exactly how the activity of the U2 splicing complex is disrupted in these cases remains uncertain, but given the lack of consensus binding sequences, no clear-cut mechanisms have been apparent. Given the finding of interactions between a mutant form of SF3B1 commonly observed in MDS cases and the germline mutation we identified, as well as the phenotypic similarity between these disorders, it is interesting to speculate about potential connections between these observations. We have not been able to

demonstrate aberrant *GATA1* mRNA splicing or expression in MDS cases in our analyses, but it is possible that we may be limited due to selection for cells lacking such an impairment. The phenotypic similarities are particularly compelling, given the observed perturbations of GATA1 protein levels that have been reported in MDS patient samples and this will be an important area for future investigation (22). Our findings also demonstrate how by reducing the overall levels of GATA1, a distinct defect in human hematopoiesis can emerge. Missense mutations in *GATA1* can result in dyserythropoiesis, thalassemia, and a variety of thrombopoietic defects (12). Lack of the full-length form of GATA1 with continued production of the short isoform can cause Diamond-Blackfan anemia (13). We have recently found that impaired translation of GATA1 in early hematopoietic progenitors underlies the more commonly observed cases of Diamond-Blackfan anemia due to ribosomal protein mutations (19). Moreover, such aberrant translation of GATA1 may also occur in other blood disorders, such as myelofibrosis (23). The results we describe in this paper extend the spectrum of these phenotypes and show how distinct alleles in a single key regulator of hematopoiesis can cause pleiotropic phenotypes, illustrating the numerous functions of GATA1 in normal human hematopoiesis.

Figure 1

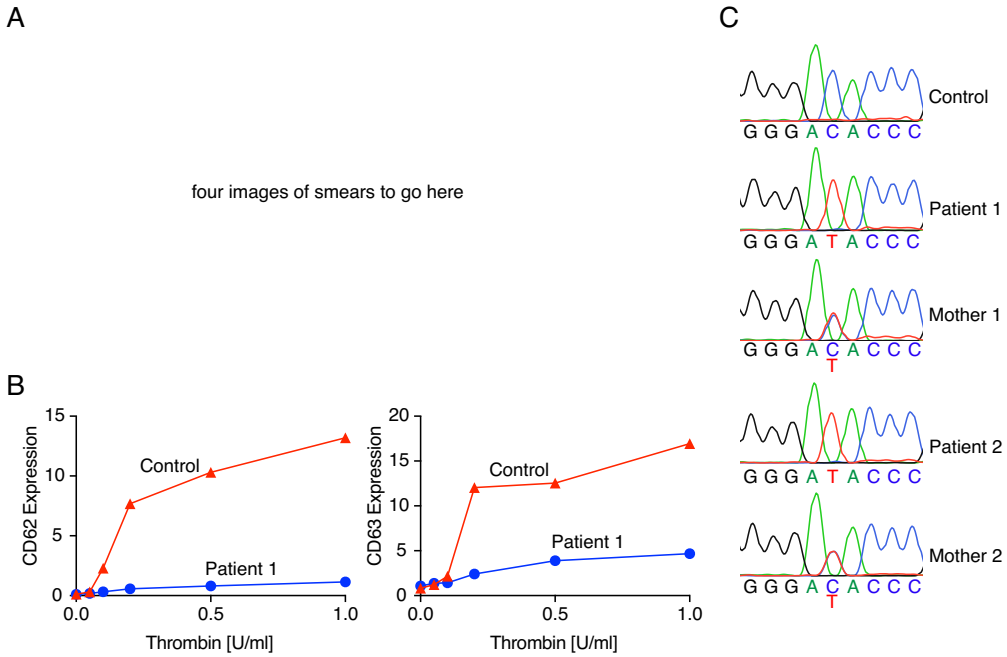


Figure 1. Identification of a *GATA1* intronic mutation in two unrelated patients with dyserythropoietic anemia. (A) Images of bone marrow aspirates from Patient 1 and 2. **(B)** The thrombin-induced expression of CD62 and CD63 on platelets from Patient 1 or a healthy control was detected by flow cytometry. There is defective expression of these antigens, indicating an α-/δ-granule deficiency. (**C**) Sequencing chromatograms of mutation in *GATA1* (ChrX:48,652,176 C>T in hg19) from a healthy control, the two patients, and their mothers.

Figure 2

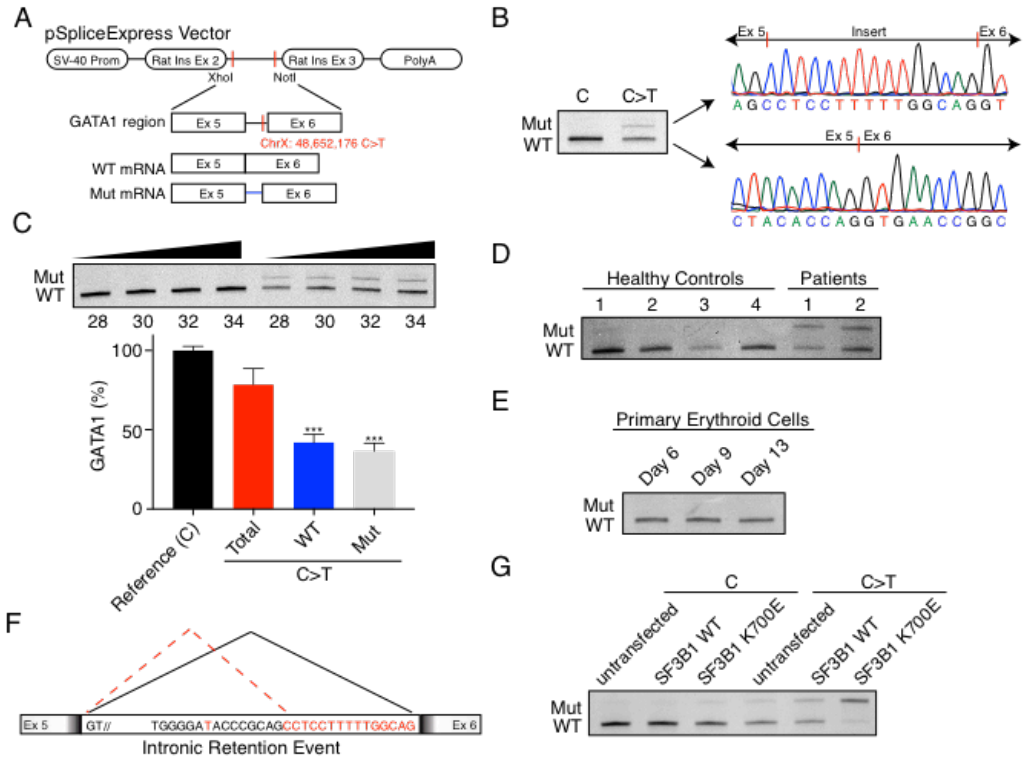


Figure 2. Decreased canonical splicing and intron retention due to a pathogenic *GATA1* mutation. (A) Schematic of the minigene assay involving exons 5 and exons 6 of *GATA1*. The minigene vector, pSpliceExpress includes two exons from Rat Insulin (Rat Ins Ex) as a control for splicing, a SV-40 promoter (Prom), and a polyadenylation site (PolyA) **(B)** Representative RT-PCR from the minigene assay in the presence or absence of the mutation shows the reduction of canonical splicing and presence of an intron retention event, which was confirmed through cloning and Sanger sequencing of the resultant products. **(C)** Semi-quantitative RT-PCR analysis of the minigene assay using 28, 30, 32, and 34 PCR cycles. The bar graph (below) depicts quantified levels of the various products, including the total amount of GATA1 (red), the amount of the wild type (WT) band (blue), or the amount of the mutant (Mut) band in the setting of the mutation involving an intron retention event (***$P< 0.001$). **(D)** RT-PCR analysis of *GATA1* exon 5 and exon 6 from control and patient peripheral blood mononuclear cells. **(E)** Representative RT-PCR analysis of *GATA1* exon 5 and exon 6 from human HSPCs undergoing erythroid differentiation. **(F)** Schematic demonstrating alternative splicing in the 5th intron of *GATA1*, including the location of the intronic mutation and the nucleotides involved in intron retention (red). The splicing patterns are depicted above with a red dashed or solid black line. **(G)** RT-PCR analysis of the minigene assay performed, as above, or with increased expression of wild type (WT) SF3B1 or the K700E SF3B1 mutated protein.

Figure 3

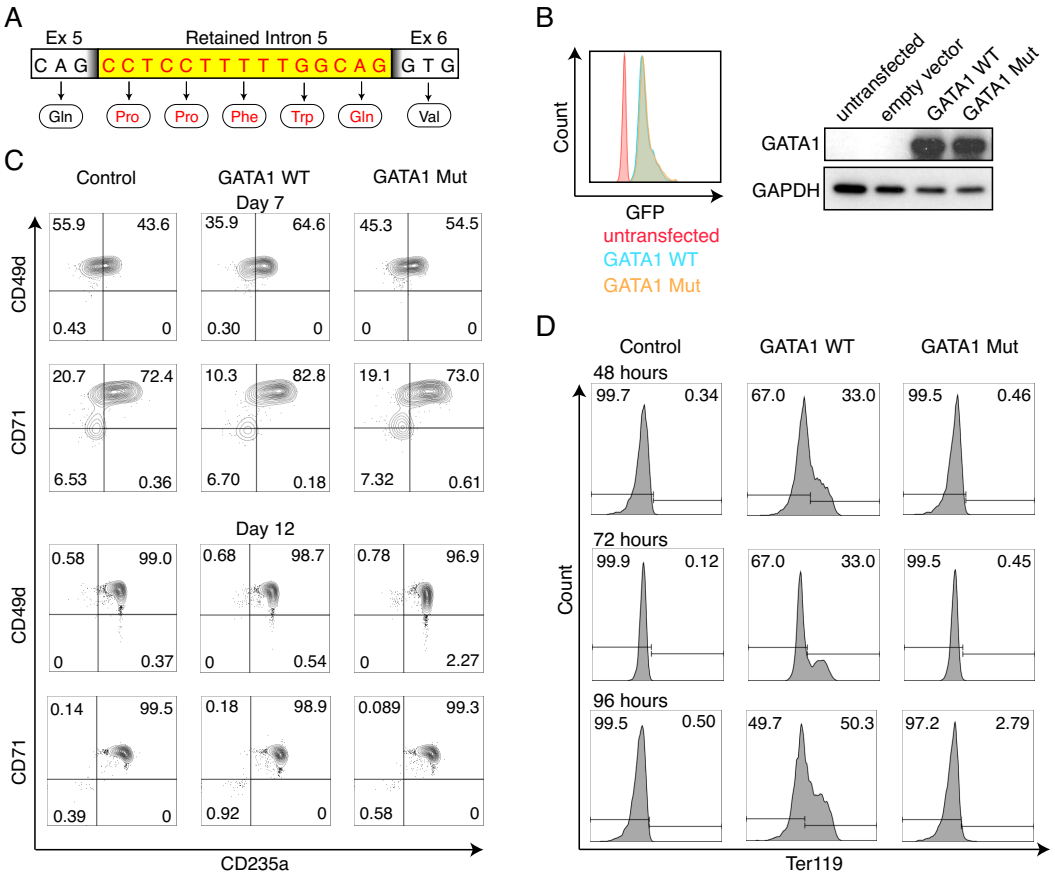


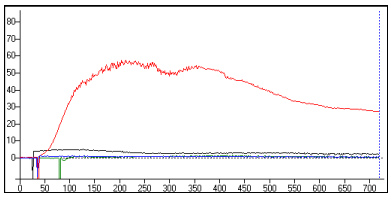
Figure 3. GATA1 variant produced through the intron retention event is expressed, but shows no function. (A) Schematic shows insertion of 5 amino acids as a result of the intron retention event. **(B)** A histogram plot (left) shows GFP levels indicating robust expression of the RNAs (GFP linked to cDNA by internal ribosomal entry site). A representative western blot of GATA1 levels from exogenous cDNA expression is show (right). **(C)** Representative flow cytometric assessment of erythroid differentiation (using CD235a, CD71, and CD49d) on day 7 and day 12 during erythroid differentiation of human HSPCs upon exogenous expression of wild type (WT) GATA1 or the mutated GATA1 (Mut) protein. **(D)** Representative histogram plots assessing the marker of mouse erythroid differentiation, Ter119, in G1E cells infected with GATA1 WT or GATA1 Mut cDNAs after 48, 72, or 96 hours.

Supplementary Figure 1

A

three additional smear images to go here

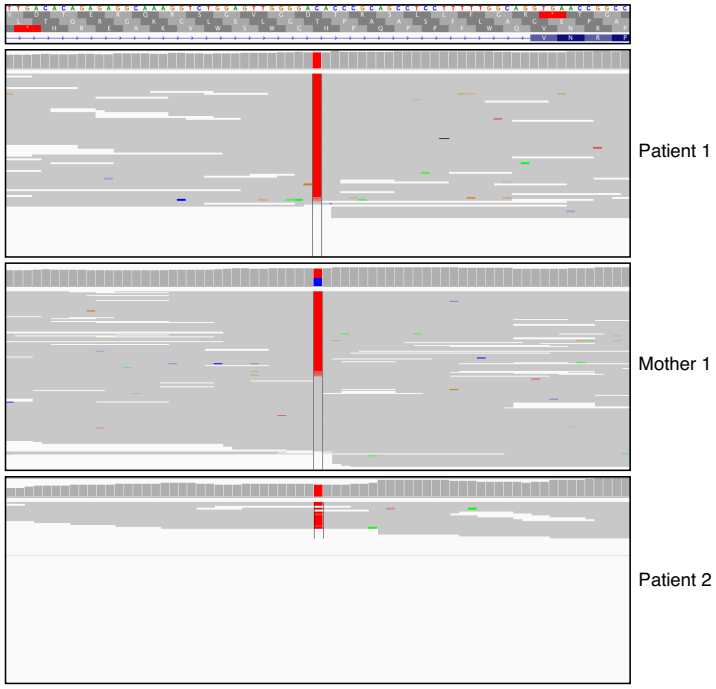
B



Agonist	Maximum Aggregation (%)	Color
Collagen [2.0 μg/ml]	5.0	black
Ristocetin [1.2 mg/ml]	56.3	red
ADP [4 μmol/l]	1.2	blue
Epinephrine [8 μmol/l]	0.7	green

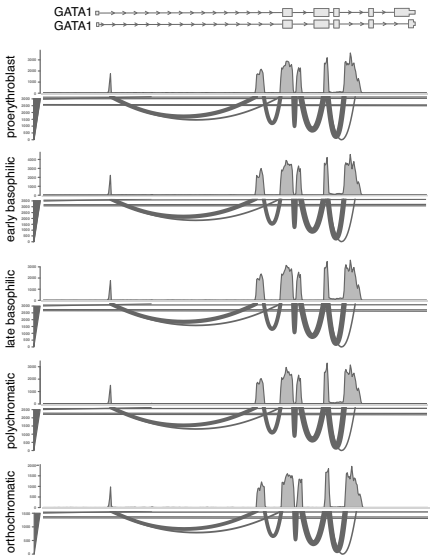
Supplemental Figure 1. Clinical phenotypes of patients with dyserythropoietic anemia and disorder hematopoiesis. (A) Histology images of bone marrow aspirates. **(B)** Platelet agglutination/aggregation analyses, as measured by light transmission, after stimulation with several agonists that are labeled in the table on the right.

Supplemental Figure 2



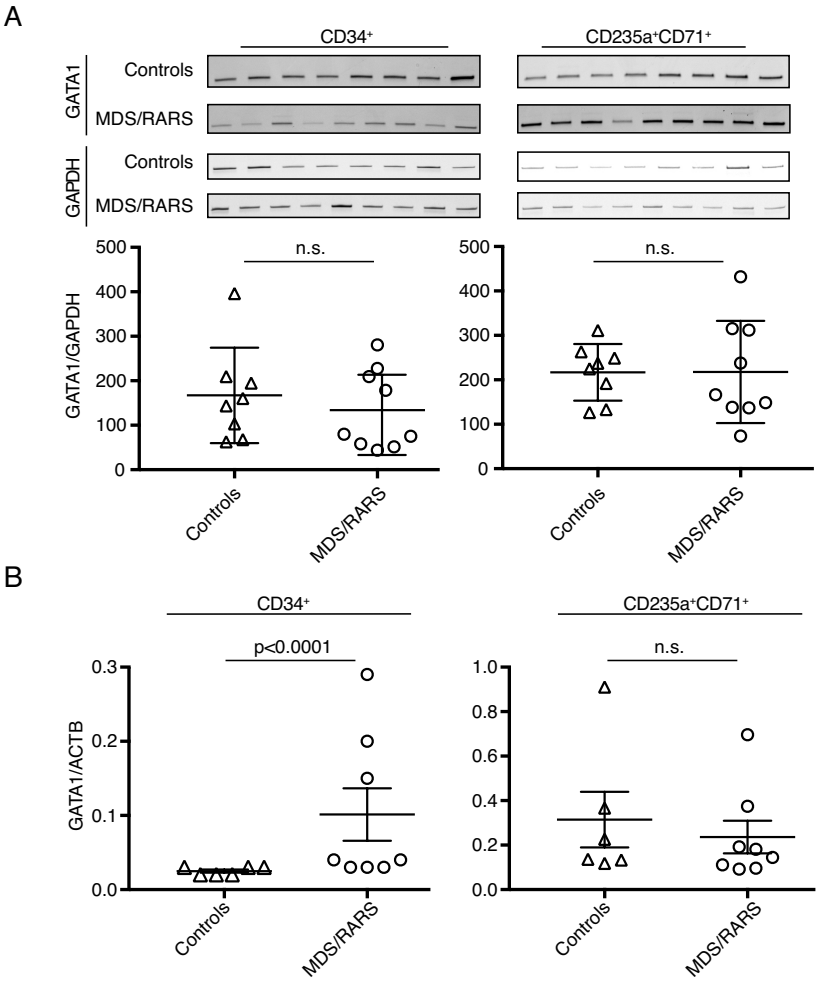
Supplemental Figure 2. Sequencing analysis of two unrelated patients with a distinct form of dyserythropoietic anemia. Integrated Genomics Viewer was used to visualize whole exome sequencing reads to verify the variant at position chrX: 48652176 (hg19 coordinates) and the sequencing coverage in this region. Patient 1 and Patient 2 are hemizygous, while Mother 1 (mother of Patient 1) is a carrier of this mutation.

Supplemental Figure 3



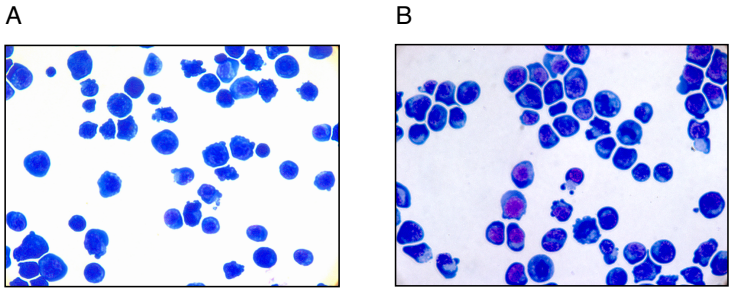
Supplemental Figure 3. RNA-seq analysis of GATA1 splicing patterns during human erythropoiesis. RNA sequencing reads in the GATA1 locus during terminal erythroid differentiation are shown. The total reads spanning each exon are shown above each axis during each stage of differentiation. A Sashimi plot of junction spanning reads is shown below for each stage. No major changes in splicing or exon usage are noted throughout erythroid differentiation.

Supplemental Figure 4



Supplemental Figure 4. Sorted bone marrow populations of *SF3B1* mutated MDS samples or controls do not have detectable differences in *GATA1* mRNA levels. (A) RT-PCR analysis of GATA1 exon 5 and exon 6 from $CD34^+$ or $CD71^+CD235a^+$ sorted bone marrow cells from *SF3B1* mutated MDS samples or healthy controls. GAPDH is used as a loading control. The top panels show gel and no appreciable levels of the intron retention event between exons 5 and 6 can be seen in the upper bands. The panels below show quantification of these gels using ImageJ. N.S. is non-significant. **(B)** Quantitative RT-PCR analysis of GATA1 exon 2 and exon 3 from the same samples above normalized to ACTB.

Supplemental Figure 5



Supplemental Figure 5. G1E cells overexpressing GATA1 protein from intron retention event demonstrate delayed maturation. (A) Cytocentrifuge analysis of GFP$^+$ sorted G1E cells overexpressing GATA1 WT cDNA, 48 hours after infection. **(B)** Cytocentrifuge analysis of GFP$^+$ sorted G1E cells overexpressing GATA1 mutant cDNA 48 hours after infection.

METHODS

Case Reports.

Patient 1 was a boy born at full term to healthy unrelated parents from Bulgaria (mother) and Togo (father). On first day of life anemia with hemoglobin (Hb) of 9.4 g/dL and concomitant jaundice with elevated lactate dehydrogenase (LDH) were noted. Besides a second-degree hypospadias, no syndromic features were present. The patient required transfusions at 6 weeks and 3 months of age because of Hb drops to the 5 g/dL range with an inadequate reticulocyte response. Subsequently he maintained a stable Hb of 8-10 g/dL with persistent red cell macrocytosis and a few giant platelets. At the age of 3.5 years, a respiratory tract infection with febrile seizure occurred with transiently worsened anemia (Hb 6.3 g/dL, reticulocytes at 18%) and a mild thrombocytopenia (133,000 cells per microliter). The patient had elevated LDH with no signs of active hemolysis. Extensive evaluations performed during and between the acute episodes for non-immune hemolytic anemias, infectious, metabolic, and autoimmune disorders, as well as bone marrow failure syndromes (including mitomycin C induced DNA damage testing and telomere length measurements), were unrevealing. An erythrocyte adenosine deaminase level was elevated at 2.79 U/g Hb. Multiple bone marrow examinations revealed normal cellularity with signs of dyserythropoiesis with macrocytic and occasional megaloblastic differentiation, mild dysplasia of the megakaryocytic lineage with hypolobated forms, and hypogranulated neutrophils. Outside of occasional incurrent illnesses, the Hb remained stable at levels of 9-10 g/dl with macrocytosis (MCV ~100 fl) and a persistently elevated fetal Hb (~20-23%), while the platelet counts remained in the low normal range. At the age of 7, when presenting with *Mycoplasma* pneumonia, there was severe epistaxis

and moderate thrombocytopenia (45,000 cells per microliter), as well as platelet function abnormalities that persisted even as the platelet counts normalized. The family medical history was unremarkable with the exception of a maternal first cousin with a chronic anemia of unclear etiology.

Patient 2: Awaiting clinical data

The whole-exome sequencing data are available in the dbGaP database (http://www.ncbi.nlm.nih.gov/gap) under the accession number phs000474.v2.p1.

Statistics. All pairwise comparisons were performed using the 2-tailed Student's *t* test, unless otherwise indicated. Differences were considered significant if the *P* value was less than 0.05.

Study approval. All family members provided written informed consent to participate in this study. The IRBs of Boston Children's Hospital, Massachusetts Institute of Technology, the University of Michigan, and the University of Freiburg approved the study protocols.

SUPPLEMENTAL METHODS

Platelet Aggregometry and Flow Cytometry

Platelet agglutination/ aggregation were analyzed using the following agonists: ristocetin (1.2 mg/ml), collagen (2.0 μg/ml), adenosine diphosphate (ADP; 4.0 μmol/l), and epinephrine (8 μmol/l), as described previously (1). For flow cytometric analyses, diluted platelet rich plasma ($5x10^7$ platelets/ml) was stimulated with different concentrations of thrombin (0.025-1.0 U/ml) in the presence of 1.25 mM Gly-Pro-Arg-Pro. Platelets were stained by monoclonal anti-CD62P (CLB-thromb/6-FITC) and anti-CD63 (CLB-gran/12-FITC) antibodies and analyzed by flow cytometry as previously described (2).

Whole Exome Sequencing & Related Genetic Analyses

The patients described in this paper are part of a rare blood disorder cohort that has been studied through the use of whole exome sequencing (WES). WES was performed as previously described (3, 4). The resultant variant call file (in hg19 coordinates) was annotated with VEP v89 (5) and rare variants (based on ExAC v0.3.1 and GnomAD r2.0.2) (6) (http://gnomad.broadinstitute.org/) were identified using a combination of the Genome Analysis Toolkit, Bcftools, and Gemini (7-9). No rare (<0.01% allele frequency in ExAC and GnomAD) loss-of-function or missense variants were identified in any known red blood cell disorder genes (ANK1, SPTB, SPTA1, SLC4A1, EPB42, EPB41, PIEZO1, KCNN4, GLUT1, G6PD, PKLR, NT5C3A, HK1, GPI, PGK1, ALDOA, TPI1, PFKM, ALAS2, FECH, UROS, CDAN1, SEC23B, KIF23, KLF1, GATA1, HBB, HBA1, HBA2, RPS7, RPS10, RPS15A, RPS17, RPS19, RPS20, RPS24, RPS26, RPS27,

RPS28, RPS29, RPS31, RPL5, RPL11, RPL15, RPL18, RPL26, RPL27, RPL35, RPL35A, TSR2, EPO), including those known to cause dyserythropoietic anemia. We next expanded our search to include all rare variants in these genes, and noted chrX: 48,652,176 C>T in the fifth intron of GATA1 harbored by these two patients, but absent from ExAC and GnomAD. The variant was hemizygous in both patients and carried in both mothers, suggesting a model of complete penetrance (the mother of Patient 1 was studied by exome sequencing, while the mother of Patient 2 had sequence analysis by Sanger sequencing). All mutations were confirmed from genomic DNA samples of the patients or family members by Sanger sequencing.

Minigene Assay and Splicing Analysis

The *GATA1* region flanking exon 5 and exon 6 was PCR amplified from genomic DNA from Patient 1 and from a healthy unrelated individual with the addition of XhoI and NotI restriction enzyme sites (Primers are the following: GATA1 intron 4: ATCATCCTCGAGTCTTGGGTCCTCCTGACATC; GATA1 intron 5: ATCATCGCGGCCGCCACATGGTCACACATTGCAG) for cloning into the pSpliceExpress vector (Addgene) (10). The constructs were transfected into human embryonic kidney (HEK) 293T cell (ATCC) with FuGENE (Promega). After reverse transcription of RNA obtained 48 hours post-transfection, RT-PCR was performed and amplified fragments were cloned using the TOPO TA Kit (Thermo Fisher) to enable confirmation of sequences through Sanger sequencing. For RT-PCR analyses, the segment of interest in *GATA1* was amplified with the following primers (GATA1 Exon 5; AGTGGGGATCCCGTGTG and GATA1 Exon 6; ATCCTTCCGCATGGTCAGT).

Amplified products were linearized with 1X TBE-Urea (BioRad), incubated for 10 minutes at 95°C, run on a 10% TBE-Urea Gel (BioRad), and subsequently visualized on a GelDoc instrument (BioRad) after 15 minutes of staining in ethidium bromide/ 1X TBE buffer. For semi-quantitative analyses, bands were quantified from scanned images using ImageJ software.

Cell Culture and Lentiviral Transduction

Human primary adult bone marrow derived $CD34^+$ hematopoietic stem and progenitor cells (HSPCs) were obtained from Fred Hutchinson Cancer Research Institute and cultured as previously described (11, 12). The wild type or mutated (involving a 5 amino acid insertion corresponding with the observed intron retention event) GATA1 cDNA constructs were cloned into the HMD vector and transfected into 293T cells for lentiviral production, along with helper plasmids as described previously (11). The primary HSPCs undergoing erythroid differentiation were infected with lentiviruses on day 2 of differentiation. Multiple donors were used in different experiments, which all yielded similar results. On day 5 of differentiation, cells were sorted for GFP and maintained in culture for further terminal erythroid differentiation (13). Differentiating cells were collected on days 7, 9 and 12 for flow cytometric analysis with erythroid markers CD235a (HIR2), CD71 (OKT9), and CD49d (9F10) using an Accuri Instrument (BD Biosciences). G1E cells were kindly provided by the laboratory of Dr. Mitchell Weiss and cultured with 15% FBS, 1-thiogycerol, murine stem cell factor (SCF, Peprotech) and human erythropoietin (EPO, Amgen). G1E cells were infected with the lentiviruses discussed above and were either subjected to flow cytometric analysis with the mouse

erythroid marker Ter119 and GFP using an Accuri cytometer (BD Biosciences) or were sorted and subject to downstream analyses. Lentiviruses expressing SF3B1 wild type or K700E mutants, as described (14), were used to infect 293T cells one day prior to transfection, as above for the mini-gene assay. Cytocentrifugation was performed with approximately 100,000-200,000 cells using a Shandon Cytospin 4 on to polylysine-coated slides at 300X RPM for 4 minutes. After drying, samples were stained with May-Grünwald for 5 minutes then with Giemsa for 15 minutes, as previously described (15).

RNA-seq Analysis

Single-end RNA-seq datasets for five distinct maturation stages of primary human erythroid cells in biological triplicates were obtained from NCBI Gene Expression Omnibus (GEO; GSE53983) (16). These datasets were aligned to the Ensembl GRCh37 r75 genome and gene annotations using 2-Pass STAR alignment. Resultant alignments were imported into R.

Analysis of MDS Patient Samples

Cryopreserved bone marrow mononuclear cells from MDS patients with refractory anemia with ringed sideroblasts and healthy controls were thawed, washed, and recovered prior to staining with fluorescently-conjugated antibodies targeting CD235a (HIR2), CD71 (OKT9), and CD34 (4H11). Flow cytometric cell sorting for $CD235a^{+}CD71^{+}$ and $CD34^{+}$ was performed using a FACSAria II Cell Sorter (BD Biosciences). Following sorting, the cells were washed and RNA was isolated using the Ambion small RNA extraction kit (Zymo). cDNA was synthesized from the recovered

RNA samples (20-40 ng) using the iScript cDNA Kit (BioRad). RT-PCR was carried out with with GATA1 primers and anormalization controls (primers GATA1 Exon 5: AGTGGGGATCCCGTGTG, GATA1 Exon 6: ATCCTTCCGCATGGTCAGT, GAPDH Exon 3: CACCAGGGCTGCTTTTAACT, and GAPDH Exon 4: GACAAGCTTCCCGTTCTCAG). Bands were visualized on denaturating gels and quantified as noted above. Samples were also analyzed with quantitative RT-PCR for GATA1 using a CFX96 Real Time Thermocycler (Biorad) with SYBR Green (BioRad) (primers GATA1 Exon 2: CCCCAGTTTGTGGATCCT and GATA1 Exon 3: CACAGTTGAGGCAGGGTAGAG, human ACTB Exon 3: AGAAAATCTGGCACCACACC, and human ACTB Exon 4: GGGGTGTTGAAGGTCTCAAA).

REFERENCES

1. Kim AR, Ulirsch JC, Wilmes S, Unal E, Moraga I, Karakukcu M, et al. Functional Selectivity in Cytokine Signaling Revealed Through a Pathogenic EPO Mutation. *Cell.* 2017;168(6):1053-64 e15.

2. Polfus LM, Khajuria RK, Schick UM, Pankratz N, Pazoki R, Brody JA, et al. Whole-Exome Sequencing Identifies Loci Associated with Blood Cell Traits and Reveals a Role for Alternative GFI1B Splice Variants in Human Hematopoiesis. *Am J Hum Genet.* 2016;99(2):481-8.

3. Basak A, Hancarova M, Ulirsch JC, Balci TB, Trkova M, Pelisek M, et al. BCL11A deletions result in fetal hemoglobin persistence and neurodevelopmental alterations. *J Clin Invest.* 2015;125(6):2363-8.

4. Sankaran VG, Weiss, MJ. Anemia: progress in molecular mechanisms and therapies. *Nat Med.* 2015;21(3):221-30.

5. Sankaran VG, Ulirsch JC, Tchaikovskii V, Ludwig LS, Wakabayashi A, Kadirvel S, et al. X-linked macrocytic dyserythropoietic anemia in females with an ALAS2 mutation. *J Clin Invest.* 2015;125(4):1665-9.

6. Wakabayashi A, Ulirsch JC, Ludwig LS, Fiorini C, Yasuda M, Choudhuri A, et al. Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. *Proc Natl Acad Sci U S A.* 2016;113(16):4434-9.

7. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, et al. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell.* 2016;165(6):1530-45.

8. Gasperini M, Findlay GM, McKenna A, Milbank JH, Lee C, Zhang MD, et al. CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *Am J Hum Genet.* 2017;101(2):192-205.

9. Rosenberg AB, Patwardhan RP, Shendure J, and Seelig G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell.* 2015;163(3):698-711.

10. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med.* 2017;9(386).

11. Sperling AS, Gibson CJ, and Ebert BL. The genetics of myelodysplastic syndrome: from clonal haematopoiesis to secondary leukaemia. *Nat Rev Cancer.* 2017;17(1):5-19.

12. Crispino JD, and Horwitz MS. GATA factor mutations in hematologic disease. *Blood.* 2017;129(15):2103-10.

13. Sankaran VG, Ghazvinian R, Do R, Thiru P, Vergilio JA, Beggs AH, et al. Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. *J Clin Invest.* 2012;122(7):2439-43.

14. Campbell AE, Wilkinson-White L, Mackay JP, Matthews JM, and Blobel GA. Analysis of disease-causing GATA1 mutations in murine gene complementation systems. *Blood.* 2013;121(26):5218-27.

15. Kishore S, Khanna A, and Stamm S. Rapid generation of splicing reporters with pSpliceExpress. *Gene.* 2008;427(1-2):104-10.

16. Hu J, Liu J, Xue F, Halverson G, Reid M, Guo A, et al. Isolation and functional characterization of human erythroblasts at distinct stages: implications for understanding of normal and disordered erythropoiesis in vivo. *Blood.* 2013;121(16):3246-53.

17. Obeng EA, Chappell RJ, Seiler M, Chen MC, Campagna DR, Schmidt PJ, et al. Physiologic Expression of Sf3b1(K700E) Causes Impaired Erythropoiesis, Aberrant Splicing, and Sensitivity to Therapeutic Spliceosome Modulation. *Cancer Cell.* 2016;30(3):404-17.

18. DeBoever C, Ghia EM, Shepard PJ, Rassenti L, Barrett CL, Jepsen K, et al. Transcriptome sequencing reveals potential mechanism of cryptic 3' splice site selection in SF3B1-mutated cancers. *PLoS Comput Biol.* 2015;11(3):e1004105.

19. Ludwig LS, Gazda HT, Eng JC, Eichhorn SW, Thiru P, Ghazvinian R, et al. Altered translation of GATA1 in Diamond-Blackfan anemia. *Nat Med.* 2014;20(7):748-53.

20. Giani FC, Fiorini C, Wakabayashi A, Ludwig LS, Salem RM, Jobaliya CD, et al. Targeted Application of Human Genetic Variation Can Improve Red Blood Cell Production from Stem Cells. *Cell Stem Cell.* 2016;18(1):73-8.

21. Rylski M, Welch JJ, Chen YY, Letting DL, Diehl JA, Chodosh LA, et al. GATA-1-mediated proliferation arrest during erythroid maturation. *Mol Cell Biol.* 2003;23(14):5031-42.

22. Frisan E VJ, de Thonel A, Pierre-Eugène C, Sternberg A, Arlet JB, Floquet C, Gyan E, Kosmider O, Dreyfus F, Gabet AS, Courtois G, Vyas P, Ribeil JA,

Zermati Y, Lacombe C, Mayeux P, Solary E, Garrido C, Hermine O, Fontenay M. Defective nuclear localization of Hsp70 is associated with dyserythropoiesis and GATA-1 cleavage in myelodysplastic syndromes. *Blood.* 2012;119(6):1532-42.

23. Gilles L, Arslan AD, Marinaccio C, Wen QJ, Arya P, McNulty M, et al. Downregulation of GATA1 drives impaired hematopoiesis in primary myelofibrosis. *J Clin Invest.* 2017;127(4):1316-20.

CHAPTER 4. CONCLUDING REMARKS & FUTURE DIRECTIONS

4.1 Using genome editing to model disease: limitations and solutions

The ease of genome sequencing has led to the discovery of genetic mutations causing a variety of human diseases. The wealth of data has now created a new primary challenge of understanding their functional and clinical significance. Genome editing emerged over the last several years as a promising approach to manipulate virtually any sequence in the genome with a high degree of target specificity. Multiple recent proof-of-principle studies have described successful application of CRISPR/Cas9 to generate models of common diseases with well-established genetic alterations, such as lung cancer with CD74-ROS1 translocation and EML4-ALK and KIF5B-RET inversions (Choi and Meyerson 2014), acute myeloid leukemia with mutations in epigenetic modifiers, transcription factors, and cytokine signaling mediators (Heckl et al., 2014), colorectal cancer with mutations in defined tumor suppressor genes and oncogenes (Matano et al., 2015), and Alzheimer's disease with mutations in amyloid precursor protein (Paquet et al., 2016). However, no study to my awareness has modeled a novel disease with a novel mutation and provided insight into disease pathophysiology.

Genome editing is the ideal approach to investigate a new disease-causing mutation because it can leave the rest of the genome unperturbed, preserving all known and unknown *cis*-acting regulatory elements and *trans*-acting factors in a cellular context that most resembles primary cells. The patients in our study presented with a distinct form of dyserythropoietic anemia characterized by persistence of fetal hemoglobin. Induction of fetal hemoglobin expression through childhood and beyond remains a topic of great interest because it can be harnessed for therapeutic purposes. A cellular model

that faithfully mimics the intronic mutation in its endogenous context provides opportunities to ask questions that exogenous expression studies are not well suited for. For example, how do global expression profiles and genome-wide occupancy sites of *GATA1* differ between wild type and mutant cells and what can we extrapolate from them about fetal hemoglobin induction? The mutant *GATA1* appears to be an inactive protein with no dominant negative activity, suggesting that the precise expression of functional *GATA1* is critical to health and disease. As expression from an exogenous construct is difficult to control, replicating a mutation through genome editing remains the "gold standard" for elucidating a precise relationship between genotype and phenotype.

In the course of creating a cellular model for the unusual dyserythropoietic anemia observed in our patients, I have discovered several key limitations of using genome editing to model a novel disease with a novel mutation. In the following sections, I will discuss these limitations and propose solutions:

The disease-causing mutation impairs survival

When a mutation and its corresponding phenotype are not well characterized, the ability of a cellular or animal model to survive is an important assumption that warrants close examination. Here, I showed that editing of the target locus in *GATA1* favors intronic sequences upstream from the Cas9 cleavage site and deletions of critical downstream sequences such as the canonical splice acceptor site and the last coding exon are exclusively observed in heterozygous clones. These results suggest that biallelic disruption of *GATA1* predicted to alter translation and function may be selected against, as K562 cells may require a sufficient level of functional GATA1 for self-renewal.

Furthermore, isogenic cell lines that recapitulated the desired C>T intronic mutation in both alleles demonstrated concurrent deletions of the pathogenic alternative splice acceptor site (Mutants 1-2) or substitution of a nearby base (Mutant 3). It is possible that these additional modifications, which are upstream of the sgRNA PAM sequence not involved in Cas9-sgRNA binding, were found because they rescue K562 cells from the defect introduced by the intronic mutation. A more definitive experiment to show that K562 cells require GATA1 for survival would be to target the last exon for editing. CRISPR/Cas9 is particularly well suited for knockout studies because indels occur frequently with NHEJ and can cause frameshift mutations. In this study, I showed that it possible to generate nonspecific indels at the target locus with up to ~80% efficiency when the guide RNA is functionally validated and nucleofection parameters are optimized. Inability to detect homozygous frameshift mutations of any length would provide strong evidence that K562 cells cannot tolerate complete loss of GATA1 function. This framework can be applied to future studies using genome editing to create a disease model when there is suspicion that the mutation jeopardizes survival. Prior to creating a disease model, one can target the gene for knockout via a validated guide RNA. If the cell or organism can survive without functional expression of the gene, then it is biologically feasible to create the disease model. However, if functional expression is required for survival, then an alternative approach such as generating patient-specific iPSCs from skin fibroblasts may be pursued. Differentiation of iPSCs can elucidate at which developmental stage the gene of interest becomes critical. Studies have described directed differentiation of iPSCs into red blood cells and a variety of other lineages (Dias et al. 2011; Wang et al., 2014; Dorn et al., 2015; Paquet et al., 2016; Pashos et al., 2017).

The mutation lacks nearby targeting sequences

Genome editing can alter any nucleotide in theory, but in practice it is constrained by specific sequence requirements. The use of CRISPR/Cas9, for example, depends on the presence of unique 20-nucleotide sequences with an adjacent PAM to minimize off-target effects and enable Cas9 binding. The canonical PAM for *S. pyogenes* Cas9 is NGG, which occurs approximately every 8 bp in the human genome (Cong et al., 2013). In the 10 nucleotides flanking my mutation of interest, computational analysis identified only 1 suitable sgRNA targeting sequence. Kwart et al. (2017) measured mutation incorporation as a function of Cas9 cut-to-mutation distance and found that there is an exponential drop as distance increases. Homozygous mutation incorporation becomes inefficient (defined as <30% of clones with HDR) at a cut-to-mutation distance of >10 bp. Of the 9 candidate sgRNAs I designed using the computational method, only one (sgRNA 5) was functional via transfection of 293T cells and surveyor nuclease assay. Unfortunately, it has a cut-to-mutation distance of 18 bp and the following experiments showed that the rate of mutation incorporation, even at overall editing efficiency of ~80%, was only 3.8% (5 out of 133 clones). My study demonstrates that sequence constraint is a real limitation of genome editing with CRISPR/Cas9, especially for single nucleotide changes. One strategy to overcome this limitation is to use engineered Cas9 nucleases with altered PAM specificities that have been validated in human cells and zebrafish (Kleinstiver et al., 2015). Alternatively, ZFNs and TALENs do not require PAM and can recognize target sequences of various lengths. Although assembling DNA binding domains is more challenging than synthesizing sgRNAs, many thousands of ZFNs and TALENs are already commercially available and validated (Gaj et al, 2013).

One study compared TALENs and CRISPR/Cas9 in targeting a single nucleotide mutation in *HBB* in β-thalassemia-derived iPSCs and found that TALENs mediated higher HDR and had fewer off-target events (Xu et al., 2015).

Homology directed repair has low efficiency

HDR is required to create a precise mutation, insertion, or deletion. Classical homologous recombination is a difficult approach to targeted gene modification due to its low efficiency. Since the discovery that induction of a DSB increases the frequency of HDR by several orders of magnitude, targeted nucleases have played an indispensable role in HDR-mediated genome editing. HDR occurs much less frequently than NHEJ and is generally significant only in dividing cells (Saleh-Gohari and Helleday 2004). Furthermore, during HDR the full repair template is not always read, and whether a mutation is incorporated depends on its distance from the Cas9 cleavage site (Kwart et al., 2017). In this study, I increased the mass ratio of ssODN repair template to sgRNA to Cas9 from 2:2:1 to 10:3:1 to promote the likelihood of HDR. Among 133 clones, I observed 8 recombination events that incorporated the mutation (3 homozygous T/T clones and 2 heterozygous T/C clones). Pashos et al. (2017) recapitulated DNA variants associated with regulation of blood lipid traits in human pluripotent stem cells using CRISPR/Cas9 and also highlighted the low efficiency of HDR. For one variant, they obtained several heterozygous clones out of 140 but no homozygous clones. For another variant, they obtained a single heterozygous clone from 672 clones. For the third variant, they were not able to obtain any heterozygous or homozygous clones after screening a large number. In this case, they changed the repair template from ssODN to a targeting

vector with 500-bp homology arms and a puromycin resistance cassette and eventually isolated several homozygous clones. Recently, many groups have described even more effective strategies to optimize HDR. Lin et al. (2014) combined cell cycle synchronization using reversible chemical inhibitors with direct nucleofection of pre-assembled Cas9 ribunucleoprotein and observed HDR rates up to 38%. Yang et al. (2016) arrested human pluripotent stem cells at the G2/M phase with anti-cancer agents ABT and nocodazole and observed a 6-fold increase in correct targeting cassette integration. Interestingly, concurrent inhibition of NHEJ with SCR7 did not further increase HDR, suggesting that HDR is the major repair mechanism following G2/M arrest. Richardson et al. (2016) rationally designed asymmetric ssODN repair template based on the DNA strand that Cas9 releases first and increased HDR rate up to 60%. Finally, Song and Stieger (2017) compared circular plasmid, linearized plasmid, and PCR products as the DNA repair template and observed highest HDR activity with linearized plasmid.

Recurrent editing creates unwanted modifications

NHEJ occurs much more frequently than HDR during most of the cell cycle. I observed that the desired C>T mutation was often accompanied by deletion of the PAM and seed sequence of sgRNA 5 upstream of the canonical splice acceptor site. This result suggests that treasured clones that have incorporated the mutation via HDR may continue to undergo repeated editing until the targeting site of Cas9-sgRNA is sufficiently obliterated. Previously, I described a novel approach to isolate the desired mutation without additional modifications using two rounds of genome editing in a framework called CORRECT (Kwart et al., 2017). First, the desired mutation and a CRISPR/Cas-

blocking mutation positioned in the PAM sequence ("re-Cas" method) or the 3' seed sequence of the target sgRNA ("re-Guide" method) are created using a repair template containing both mutations. Then, the CRISPR/Cas-blocking mutation is removed using VRER-Cas9 ("re-Cas"), an engineered Cas9 that recognizes the previously altered PAM sequence, or a new sgRNA ("re-Guide") that recognizes the previously altered 3' seed sequence, with a "correct" repair template containing only the desired mutation. The team showed introducing pathogenic mutations with silent CRISPR/Cas-blocking mutations increased HDR accuracy dramatically. They then applied this approach to generate human iPSCs with mutations in amyloid precursor protein (APP^{Swe}) and presenilin 1 ($PSEN1^{M146V}$) known to cause early onset Alzheimer's disease (Paquet et al., 2016). They derived cortical neurons from edited iPSCs and observed genotype-dependent disease-associated phenotype, such as increased total amyloid-β levels and increased ratio of the 42-residue to 40-residue Aβ peptide. Mutant 3 in my study has the desired homozygous C>T mutation but developed a *de novo* homozygous C>A mutation 10 bp downstream, which is also 8 bp upstream from the Cas9 cleavage site and 2 bp upstream from the PAM sequence. If K562 cells can survive with the homozygous C>T mutation and no additional modifications, then this mutant would be an ideal candidate for further rounds of genome editing via CORRECT to reverse the C>A mutation, which is already <10 bp from the Cas9 cut site, an ideal distance for HDR.

Guide RNA has off-target effects

The main strength of using genome editing to model disease is the ability to recapitulate mutations on an isogenic background and minimize potential confounders.

All 9 guide RNAs I designed for *GATA1* editing had computed scores ≥ 50 and were deemed high quality based on faithfulness of on-target activity. However, the estimated numbers of off-target sites ranged from 78 (sgRNA 5) to 478 (sgRNA 9) and raise the question of how often off-target genome editing occurs and whether observed phenotypic changes in a cellular model can truly be attributed to the mutation of interest. Recently, two groups developed *in vitro* biochemical methods for genome-wide assessment of CRISPR/Cas9 off-target sites. Tsai et al. (2017) used circularization to report cleavage effects in a technique called CIRCLE-seq. Genomic DNA is sheared and circularized, after which cleavage reactions are performed with Cas9 and *in vitro* transcribed sgRNA. Circular DNA molecules containing targeted sequences are linearized for adaptor ligation, PCR amplification, and high-throughput sequencing while unperturbed DNA remain circularized. Using genomic DNA from K562 cells, the authors identified 182 off-target sites associated with an sgRNA targeting the human *HBB* gene. Cameron et al. (2017) used biotin ligation to report cleavage effects in another technique called SITE-seq. Genomic DNA is digested with Cas9, after which cleaved DNA ends are tagged with biotin for subsequent enrichment and sequencing. Studies have described multiple strategies to minimize off-target effects. First, alteration of sgRNA including 3' (tracrRNA) truncation, 5' end shortening, and addition of GG to 5' end can significantly improve target specificity and reduce off-target mutagenesis (Zhang et al, 2015). Second, decreasing the amount of sgRNA and Cas9 delivered can increase specificity but also lead to decreased on-target cleavage. Third, the D10 mutant nickase version of Cas9 cleaves only one strand and can be with paired two sgRNAs, a strategy which can reduce off-target activity by 50-1,500 fold (Ran et al., 2013). Finally, catalytically inactive Cas9

fused to *Fok*I nuclease domain (fCas9) has >140-fold higher target specificity than wild-type Cas9 and at least 4-fold that of paired nickases (Guilinger et al., 2014; Tsai et al, 2014). Since multiple clones are unlikely to have the same off-target effects, consistent phenotypic differences between multiple wild type and edited clones would argue for a true effect from the intended mutation.

4.2 Alternative approaches can efficiently assess function of genetic variants

Recent progress has overcome many barriers to creating a faithful disease model. While genome editing remains an ideal approach to elucidate the genome-phenotype relationship by replicating mutations in their endogenous context and minimizing confounders, alternative approaches can efficiently assess function when generating a disease model requires substantial troubleshooting. With the advent of next-generation sequencing, many intronic or synonymous mutations that do not appear to alter protein coding have emerged as potential causes of human disease. These mutations likely reside in *cis*-acting elements that modulate pre-mRNA splicing.

Transient expression of minigene is a classical approach to characterize alternative splicing. As the primary elements regulating alternative splicing are typically located within 200-300 nucleotides of an exon and rarely beyond the upstream and the downstream exons, a “minigene” consisting of the genomic fragment of interest can be constructed in a plasmid vector and expressed exogenously to investigate splicing (Cooper 2005). RT-PCR enables quantification of mRNA including (or excluding) a variable region. Kishore et al. (2008) made the process even more efficient by developing a new vector called pSpliceExpress, which incorporates minigenes based on site-specific

recombination between bacteriophage lambda and *E. coli* chromosome, a process commercially popularized as Gateway cloning. Using this method, one can PCR amplify the genomic fragment containing the mutation of interest and flanking exons with attachment sites and directly insert it into pSpliceExpress by *in vitro* recombination. The vector contains 2 constitutive rat insulin exons flanking the insert for positive control and a ccdB element that is excised upon recombination, facilitating selection of bacteria transformed with plasmids containing the minigene. The authors report that generation of minigenes takes less than one week. In contrast, Kwart et al. (2017) estimates that generation of a scarless isogenic cell line with the mutation of interest using genome editing in the CORRECT framework takes approximately 3 months. The minigene assay has a few potential drawbacks: different cell lines may exhibit different splicing efficiencies and the same line may exhibit changes in splicing patterns over time. Encouragingly, my colleague Nour Abdulhay showed that altered GATA1 splicing due to activation of the alternative splice acceptor site is similarly observed in our patient samples, validating transient expression of minigene as an effective approach to studying mutations that regulate splicing.

Once aberrant gene expression is confirmed, alternative approaches to establish protein function include exogenous expression of mutated cDNA in cells and complementation in an animal model null for the gene of interest. With many tools ranging from traditional approaches to newest breakthroughs in genome editing, converting our wealth of data on the genetic basis of disease to functionally and clinically useful knowledge appears to be a matter of time.

REFERENCES

Alzheimer's Disease Collaborative Group (1995). The structure of the presenilin 1 (S182) gene and identification of six novel mutations in early onset AD families. Nat. Genet. *11*, 219–222.

Cameron, P., Fuller, C.K., Donohoue, P.D., Jones, B.N., Thompson, M.S., Carter, M.M., Gradia, S., Vidal, B., Garner, E., Slorach, E.M., et al. (2017). Mapping the genomic landscape of CRISPR–Cas9 cleavage. Nature Methods *14*, 600–606.

Choi, P.S., and Meyerson, M. (2014). Targeted genomic rearrangements using CRISPR/Cas technology. Nature Communications *5*.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Cooper, T.A. (2005). Use of minigene systems to dissect alternative splicing elements. Methods *37*, 331–340.

Dias, J., Gumenyuk, M., Kang, H., Vodyanik, M., Yu, J., Thomson, J.A., and Slukvin, I.I. (2011). Generation of Red Blood Cells from Human Induced Pluripotent Stem Cells. Stem Cells and Development *20*, 1639–1647.

Dorn, I., Klich, K., Arauzo-Bravo, M.J., Radstaak, M., Santourlidis, S., Ghanjati, F., Radke, T.F., Psathaki, O.E., Hargus, G., Kramer, J., et al. (2015). Erythroid differentiation of human induced pluripotent stem cells is independent of donor cell type of origin. Haematologica *100*, 32–41.

Gaj, T., Gersbach, C.A., and Barbas, C.F. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends in Biotechnology *31*, 397–405.

Guilinger, J.P., Thompson, D.B., and Liu, D.R. (2014). Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. Nat. Biotechnol. *32*, 577–582.

Haass, C., Lemere, C.A., Capell, A., Citron, M., Seubert, P., Schenk, D., Lannfelt, L., and Selkoe, D.J. (1995). The Swedish mutation causes early-onset Alzheimer's disease by beta-secretase cleavage within the secretory pathway. Nat. Med. *1*, 1291–1296.

Heckl, D., Kowalczyk, M.S., Yudovich, D., Belizaire, R., Puram, R.V., McConkey, M.E., Thielke, A., Aster, J.C., Regev, A., and Ebert, B.L. (2014). Generation of mouse models of myeloid malignancy with combinatorial genetic lesions using CRISPR-Cas9 genome editing. Nature Biotechnology *32*, 941–946.

Kishore, S., Khanna, A., and Stamm, S. (2008). Rapid generation of splicing reporters with pSpliceExpress. Gene *427*, 104–110.

Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Topkar, V.V., Nguyen, N.T., Zheng, Z., Gonzales, A.P.W., Li, Z., Peterson, R.T., Yeh, J.-R.J., et al. (2015). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature *523*, 481–485.

Lin, S., Staahl, B.T., Alla, R.K., and Doudna, J.A. (2014). Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. Elife *3*, e04766.

Matano, M., Date, S., Shimokawa, M., Takano, A., Fujii, M., Ohta, Y., Watanabe, T., Kanai, T., and Sato, T. (2015). Modeling colorectal cancer using CRISPR-Cas9–mediated engineering of human intestinal organoids. Nature Medicine *21*, 256–262.

Paquet, D., Kwart, D., Chen, A., Sproul, A., Jacob, S., Teo, S., Olsen, K.M., Gregg, A., Noggle, S., and Tessier-Lavigne, M. (2016). Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. Nature *533*, 125–129.

Pashos, E.E., Park, Y., Wang, X., Raghavan, A., Yang, W., Abbey, D., Peters, D.T., Arbelaez, J., Hernandez, M., Kuperwasser, N., et al. (2017). Large, Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like Cells Reveal Functional Genetic Variation at Blood Lipid-Associated Loci. Cell Stem Cell *20*, 558–570.e10.

Ran, F.A., Hsu, P.D., Lin, C.-Y., Gootenberg, J.S., Konermann, S., Trevino, A.E., Scott, D.A., Inoue, A., Matoba, S., Zhang, Y., et al. (2013). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. Cell *154*, 1380–1389.

Richardson, C.D., Ray, G.J., DeWitt, M.A., Curie, G.L., and Corn, J.E. (2016). Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. Nat. Biotechnol. *34*, 339–344.

Saleh-Gohari, N., and Helleday, T. (2004). Conservative homologous recombination preferentially repairs DNA double-strand breaks in the S phase of the cell cycle in human cells. Nucleic Acids Res. *32*, 3683–3688.

Song, F., and Stieger, K. (2017). Optimizing the DNA Donor Template for Homology-Directed Repair of Double-Strand Breaks. Molecular Therapy - Nucleic Acids *7*, 53–60.

Tsai, S.Q., Wyvekens, N., Khayter, C., Foden, J.A., Thapar, V., Reyon, D., Goodwin, M.J., Aryee, M.J., and Joung, J.K. (2014). Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. Nat. Biotechnol. *32*, 569–576.

Tsai, S.Q., Nguyen, N.T., Malagon-Lopez, J., Topkar, V.V., Aryee, M.J., and Joung, J.K. (2017). CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. Nature Methods *14*, 607–614.

Wang, G., McCain, M.L., Yang, L., He, A., Pasqualini, F.S., Agarwal, A., Yuan, H., Jiang, D., Zhang, D., Zangi, L., et al. (2014). Modeling the mitochondrial cardiomyopathy of Barth syndrome with induced pluripotent stem cell and heart-on-chip technologies. Nat. Med. *20*, 616–623.

Xu, P., Tong, Y., Liu, X., Wang, T., Cheng, L., Wang, B., Lv, X., Huang, Y., and Liu, D. (2015). Both TALENs and CRISPR/Cas9 directly target the HBB IVS2–654 (C > T) mutation in β-thalassemia-derived iPSCs. Scientific Reports *5*.

Yang, D., Scavuzzo, M.A., Chmielowiec, J., Sharp, R., Bajic, A., and Borowiak, M. (2016). Enrichment of G2/M cell cycle phase in human pluripotent stem cells enhances HDR-mediated gene repair with customizable endonucleases. Sci Rep *6*, 21264.

Zhang, X.-H., Tee, L.Y., Wang, X.-G., Huang, Q.-S., and Yang, S.-H. (2015). Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. Molecular Therapy - Nucleic Acids *4*, e264.